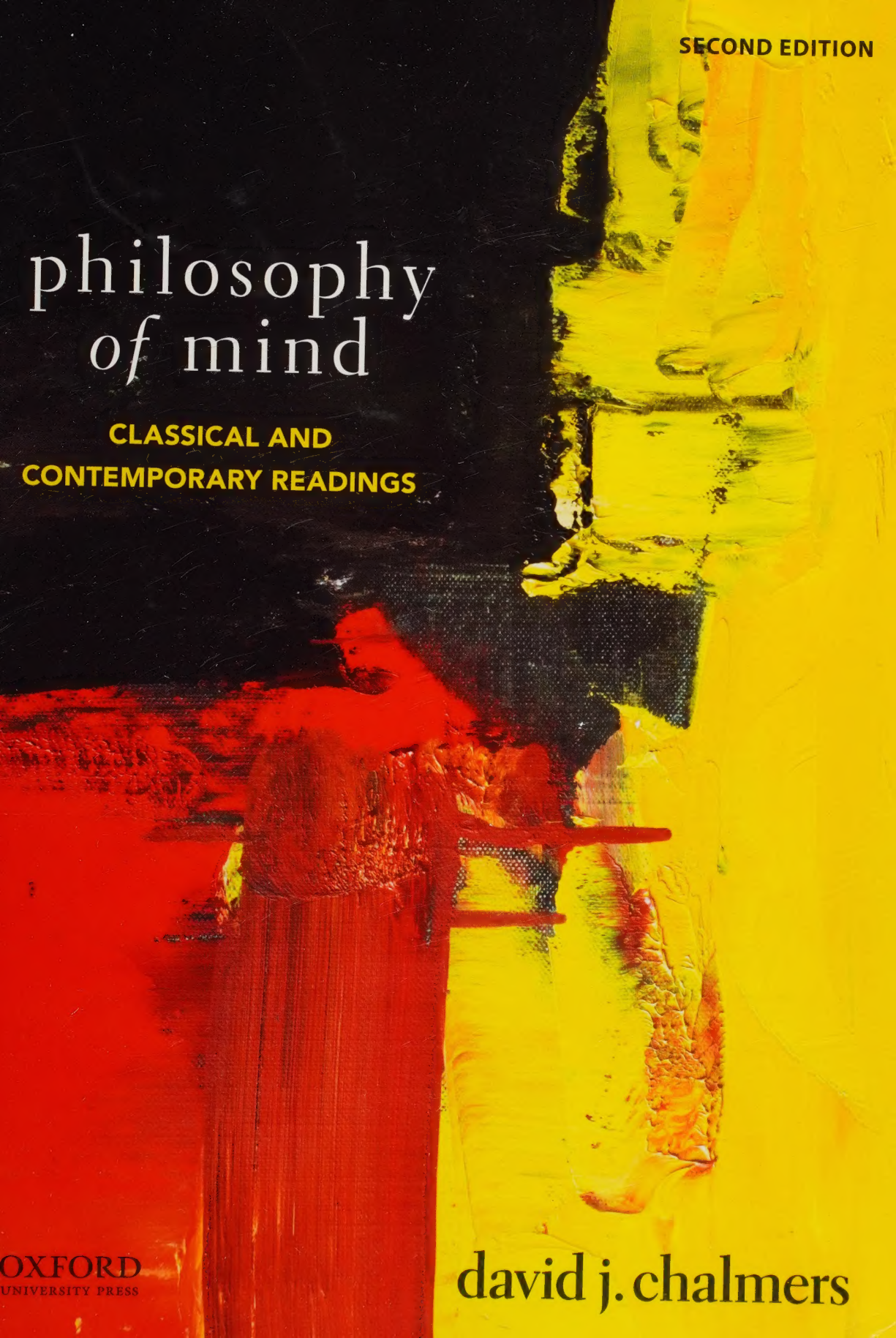
SECOND EDITION

philosophy *of* mind

CLASSICAL AND CONTEMPORARY READINGS



david j. chalmers

OXFORD
UNIVERSITY PRESS

PHILOSOPHY OF MIND

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

ISBN 9780190640859

The CIP data is on file at the Library of Congress.

Printing number: 9 8 7 6 5 4 3 2 1
Printed by LSC Communications, United States of America

CONTENTS

PREFACE

What is the mind? What is the relationship between mind and body? Is the mind the same as the brain? How can the mind affect the physical world? What is consciousness? Could a purely physical system be conscious? Can we explain subjective experience in objective terms? How does the mind represent the world? What is the nature of belief and desire? What is the relationship between consciousness and representation? Is the mind in the head or in the environment? How can we perceive the external world? How do we know our own minds? What can we know about other minds, in humans and in nonhuman animals? What is the self? What is personal identity? Can machines have minds?

These are some of the central questions in the philosophy of mind. This book is a collection of articles addressing them. If the book has a thematic focus, it is on the many aspects of the mind-body problem. What is the relationship between mind, brain, and body, and between the mental and the physical? This is perhaps the central problem in the philosophy of mind. It ramifies into any number of different questions concerning different aspects of this relationship and concerning different aspects of the mind. The articles in this book address these questions from many different angles.

This collection has seven main parts. *Foundations*: what is the nature of the mind, and what is the relationship between the mental and the physical? *Consciousness*: what is the place of consciousness in the natural world, is consciousness a physical process, and how can consciousness be understood? *Content*: how can the mind represent the world, what is the nature of thought, and are the contents of our thoughts determined by the brain or by the environment? *Perception*: what is the nature of perceptual experience, and what is the relation between perception and the external world? *Self-Knowledge and Other Minds*: how do we know about other minds, and how do we know our own minds? *The Self*: what is the self, what is the relationship between self and brain, and what is the nature of personal identity? *Artificial Intelligence*: can machines have minds, can they surpass us in intelligence, and what happens if they do?

I have written introductions to each of the parts, giving relevant background for the material in those parts, and giving pointers for further reading.

The collection includes both classical articles that make up much of the standard history of the field and contemporary articles that represent recent directions in the area. Much of the classical background to recent debates, from Descartes' dualism to various 20th-century forms of materialism, can be found in the first section of the book. The second and third sections concentrate largely on material from the past few decades, with a good representation of material at the leading edge of current research. The book contains a combination of highly accessible articles and more sophisticated pieces, so it should be suitable for use in undergraduate and graduate courses at all levels. I hope that the book will also be interesting to general readers interested in these issues.

In the almost two decades since the first edition of this book was published in 2002, the philosophy of mind has expanded in many directions. As a result, this volume has expanded too. The second edition includes new parts on perception, epistemology of mind, the self, and artificial intelligence. All of these topics have been particularly active in recent years. In the earlier parts of the book, readings have been overhauled and updated throughout.

The second edition also aims for coverage of a wider range of philosophical traditions. Where the first edition focused almost exclusively on 20th-century Anglocentric analytic philosophy of mind and some historical predecessors in Western philosophy, the second edition includes articles drawn from African philosophy, Indian philosophy, and Islamic

philosophy, as well as from more recent traditions in experimental philosophy, feminist philosophy, and phenomenology. Of course the book remains in large part a reflection of a certain tradition, and there remain many omissions.

Acknowledgments for the second edition: I would like to thank Robert Miller and Andy Blitzer at Oxford University Press for their editorial help, and Kelvin McQueen and Jonathan Simon for their help in assembling the volume. Thanks to production editor Claudia Dukeshire, copy editor Robin Reid, and proofreader Patti Brecht. I would also like to thank the following reviewers of the first edition: Torin Alter, Andrew Bailey, Bill Brewer, Curtis Brown, Alex Byrne, Elijah Chudnoff, Joe Cruz, Steven Horst, Frank Jackson, Amy Kind, Uriah Kriegel, Michelle Montague, Barbara Montero, Sarah Paul, Thomas Polger, Peter Ross, Warren Schmaus, Tim Schroeder, Eric Schwitzgebel, John Turri, and Daniel A. Weiskopf.

Acknowledgments for the first edition: I would like to thank Robert Miller of Oxford University Press for inviting me to put this book together and for all his help on the editorial front. Brad Thompson was a great help in chasing down permissions and in preparing the manuscript. Thanks are due also to Fiona Cowie, George Graham, Jaegwon Kim, and two other reviewers for their helpful comments on the contents and organization of the book. Finally, I owe a debt to the editors of previous anthologies in the area—especially Ned Block (*Readings in the Philosophy of Psychology*), David Rosenthal (*The Nature of Mind*), and William Lycan (*Mind and Cognition*)—for their example.

PHILOSOPHY OF MIND

1 Foundations

The articles in this first part of the book address foundational questions about the nature of the mind and about the relationship between the mental and the physical. Many of these questions concern the nature of mental states: states such as seeing the color red, feeling pain, experiencing anger, and desiring happiness. What is the nature of a mental state? And how are mental states related to physical states, such as states of the brain, of one's body and behavior, and of the physical world more generally?

Traditionally, views on these issues can be divided into two main classes. *Dualist* views hold that the mind is quite distinct from the body and the brain (although they may be associated in some fashion), and/or that mental states are fundamentally distinct from physical states. *Materialist* views hold that the mind is itself a broadly physical entity, and/or that mental states are fundamentally derivative on physical states. There also exist *idealist* views, on which physical states are fundamentally derivative on mental states, but these will not be as central here. The papers in this section discuss many varieties of dualist and materialist views, as well as other foundational questions about the relations between the mental and the physical.

A. Dualism

Dualist views come in two main varieties. *Interactionism* holds that the mental and physical are fundamentally distinct, but interact in both directions: physical states affect mental states, and mental states affect physical states. *Epiphenomenalism* holds that the mental and physical are fundamentally distinct and that physical states affect mental states, while denying that mental states affect physical states.

In the history of European philosophy, the most important dualist view is the interactionism of René Descartes. Descartes' most important work was his *Meditations on First Philosophy*. This is a series of six meditations, the second and sixth of which are reproduced here as chapter 1. In the first meditation, Descartes attempts to cast doubt on all of his beliefs and found that he cannot be certain that the external world exists. In the second meditation, Descartes finds that there is one thing he can be certain of: his own mind, and so his own existence ("I think, therefore I am"). He concludes that he is fundamentally "a thing that thinks." In the third through fifth meditations, Descartes infers the existence of God and uses this to justify his belief in the external world (since God would not deceive him). In the sixth meditation, Descartes reflects on the differences between the mental and the physical, and concludes that they are fundamentally distinct. He uses a number of arguments here: one can be certain about the mental but

not about the physical; the mind is indivisible while any physical entity is divisible; and most famously, one can imagine oneself existing without a body, so one must be distinct from one's body and likewise from any physical entity.

The *Meditations* argue for the distinctness of mind and body, but do not say much about the relationship. This question is addressed in more depth in Descartes' *Passions of the Soul* (chapter 2), which discusses the interaction between mind and body. Humans have a rational soul, which receives perceptions as "passions" from the brain and which performs actions through acts of will that affect the brain. Signals are passed between the brain and the soul via the pineal gland (a small gland centrally located in the brain). On this picture, mind and body involve separate substances but interact in both directions.

Princess Elisabeth of Bohemia (chapter 3) carried on an extensive correspondence with Descartes about philosophical issues. It was Elisabeth who pointed out what has become by far the most notorious problem for Cartesian dualism. This is the *interaction problem*: How could a nonphysical mind have any effects on a physical body? Elisabeth famously says, "It would be easier for me to concede matter and extension to the soul than to concede the capacity to move a body and to be moved by it to an immaterial thing." Because Descartes' mind is outside space, it is not easy to see how it could have any effect on the body, which is in space.

Descartes' ideas about the mind are widely rejected today, in large part because of the interaction problem. The idea that the pineal gland mediates this interaction has long since been rejected on physiological grounds, and no more plausible causal nexus in neurobiology has been found. Further, it is widely believed that this interaction cannot be reconciled with physics, which postulates a closed network of physical interactions, with no room for a nonphysical mind to play any role.

One can find dualist views of mind and body in many philosophical traditions. Kwame Gyekye (chapter 4) argues that the Akan tradition in Ghana is dualistic. In the Akan view, a person is constituted by *Okra* (soul) and *Sunsum* (spirit), which are immaterial, and by *Honam* (body), which is material. According to Gyekye, the Akans are interactionists: they think that *Okra* and *Sunsum* are immaterial substances located in the head that manifest themselves through their effects on body and behavior. Traditional Akan medicine relies heavily on both physical and spiritual methods, reflecting this underlying dualism.

Islamic philosophy also has a strong core of dualism about mind and body. One of the most famous arguments in the Islamic tradition is the "floating man" argument put forward by Avicenna (or Ibn Sina) in "On the Soul" (chapter 5). Avicenna imagines a man who is created floating in the air, without any feeling of his body. He suggests that the floating man will be conscious of himself and aware of his own existence, even though he is not conscious of his own body. This suggests that self-consciousness cannot be reduced to bodily consciousness. More strongly, Avicenna uses the argument to suggest that the self or soul is independent of the body. After all, the floating man is aware of himself but not aware of his body.

In reaction to the interaction problem for Cartesian varieties of dualism, some have embraced epiphenomenalism, retaining the distinctness of mind and body, but denying any causal role for mind in the physical world. Such a view is put forward by Thomas Huxley (chapter 6). Huxley addresses Descartes' view that nonhuman animals are mere automata, whose behavior is controlled entirely by their brains and who lack minds altogether. Huxley suggests that contemporary evidence favors the first aspect of this view, but does not favor the second: that is, animals' behavior is controlled entirely by their brains, but they have minds nevertheless. On this view, the mind is a sort of byproduct of the brain that has no effect on it. At the end of his article, Huxley suggests that the same goes for humans.

Epiphenomenalism has the advantage of being easier to reconcile with science than interactionist dualism, but it has the disadvantage of running strongly counter to common sense. Intuitively, it is hard to accept that our thoughts and feelings have no

effect on our behavior. Another problem is raised by Raymond Smullyan in his short fable "An Unfortunate Dualist," (chapter 7): if mind has no effect on behavior, then it has no effect on what we say about the mind, so it seems that one could remove the mind and we would go on talking about it just the same. Smullyan raises the issue as a problem for dualism in general, but it is particularly pressing for an epiphenomenalist.

FURTHER READING

The full correspondence between Elisabeth of Bohemia and Descartes can be found in Shapiro 2007. An alternative view of the Akan conception of mind and personhood is presented by Wiredu 1992. Objections to interactionism and epiphenomenalism are discussed in more detail by Kim (chapter 20) and Chalmers (chapter 29). Chalmers gives a limited defense of both interactionism and epiphenomenalism, while Jackson (chapter 30) defends an epiphenomenalist view. Elsewhere, interactionist views are defended by Foster 1991, Hodgson 1991, Popper and Eccles 1977, and Swinburne 2019, while an epiphenomenalist view is defended by Robinson 2018. A relevant collection is Smythies and Beloff 1989.

Foster, J. *The Immaterial Self: A Defense of the Cartesian Dualist Conception of Mind* (Routledge, 1991).

Hodgson, D. *The Mind Matters: Consciousness and Choice in a Quantum World* (Oxford University Press, 1991).

Popper, K., and Eccles, J. *The Self and Its Brain: An Argument for Interactionism* (Springer, 1977).

Robinson, W. S. *Epiphenomenal Mind: An Integrated Outlook on Sensations, Beliefs, and Pleasure* (Routledge, 1988).

Smythies, J. R., and Beloff, J., eds., *The Case for Dualism* (Charlottesville: University of Virginia Press, 1989).

Swinburne, R. *Are We Bodies or Souls?* (Oxford University Press. 2019).

Wiredu, K. "The African concept of personhood," in *African-American Perspectives on Biomedical Ethics*, H. Flack and E. Pellegrino, eds., (Washington, DC: Georgetown University Press, 1992).

B. Behaviorism

In the second half of the twentieth century, dualism was widely rejected, and many different forms of materialism were explored. This was both a reaction to the problems of dualism and a product of the success of physical explanations in many different domains.

Gilbert Ryle's 1949 book *The Concept of Mind* is recognizably the antecedent of much recent work in the philosophy of mind. This book argues against dualist views and puts forward a positive view of its own. Included here is the first chapter of the book, "Descartes' Myth" (chapter 8). As the title suggests, this chapter is largely a polemic against the dualism of Descartes. He accuses Descartes and others of subscribing to the "dogma of the ghost in the machine," and suggests that these views rest on a "category mistake" in posing questions about the relationship between mind and body. The mind is not to be seen as something distinct from the body and steering it from the inside, but as an aspect of the body's own activities.

Ryle's positive views (developed in other chapters of his book) are subtle and hard to summarize, but a central strand in these views involves a sort of *behaviorism*: roughly, the view that the mind is an aspect of behavior. On this view, to be in a given mental state (such as pain) is to be in a certain behavioral state (such as wincing), or at least to have a disposition to behave in certain ways (such as the disposition to express pain

if queried). Thus the mind is seen as a public aspect of human activity, rather than as a private inner aspect.

This sort of behaviorism is more explicit in Carl Hempel's "The Logical Analysis of Psychology" (chapter 9). Hempel was influenced by the logical empiricist tradition in philosophy, which held roughly that all meaningful claims can be translated into claims about observable, verifiable phenomena. In the case of the mind, this comes to the claim that meaningful claims about the mind can be translated into claims about behavior. This is a form of *logical* behaviorism, holding ultimately that what we *mean* when we make claims about the mind involves underlying claims about behavior. (This differs from scientific behaviorism, which holds roughly that the scientific study of the mind is the study of behavior.) Given that behavior itself seems to be a physical phenomenon, behaviorism can be seen as a form of materialism.

Like dualism, behaviorism has been subject to a number of objections. It seems more intuitive to say that the mind is an inner cause of behavior, rather than an aspect of behavior itself. More concretely, one can argue that any given mental state is distinct from any given behavioral state or behavioral disposition. This sort of argument is mounted by Hilary Putnam in "Brains and Behavior" (chapter 10). Putnam argues that specially trained people ("super-Spartans") might feel pain while having no associated behavioral dispositions at all. Likewise, it can be argued that a perfect actor might have any given behavioral disposition without the associated mental state. If so, mental states cannot be behavioral dispositions.

FURTHER READING

Apart from being the classic statement of a broadly behaviorist view, Ryle 1949 contains nuanced discussions of many aspects of the mind and mentality. The views of Carnap 1934, Wittgenstein 1953, and Quine 1960 also have some affinity with behaviorism. Dennett (this volume, 1987), who was a student of Ryle's, puts forward a view that can be seen as a sophisticated contemporary descendant of behaviorism. Important objections to behaviorism are given by Geach 1957 and Block 1981. Scientific behaviorism is advocated by Watson 1930 and Skinner 1971.

Block, N., "Psychologism and behaviorism," *Philosophical Review* 90 (1981): pp. 5–43.

Carnap, R., "Psychologie in physikalischer Sprache," *Erkenntnis* 3 (1932): pp. 162–76. Translated as "Psychology in physical language" in *Logical Positivism*, A. J. Ayer, ed., (Free Press, 1959).

Dennett, D. C. *The Intentional Stance* (MIT Press, 1987).

Geach, P. *Mental Acts* (Routledge & Kegan Paul, 1957).

Quine, W. V. *Word and Object* (MIT Press, 1960).

Ryle, G. *The Concept of Mind* (Hutchinson and Co., 1949).

Skinner, B. F. *Beyond Freedom and Dignity* (Alfred A. Knopf, 1971).

Watson, J. *Behaviorism* (Norton, 1930).

Wittgenstein, L. *Philosophical Investigations* (Blackwell, 1953).

C. The Identity Theory and Functionalism

The mind-brain identity theory holds that mental states are brain states. Most theorists accept that mental states are at least associated or correlate with brain states: for example, feeling pain might be correlated with a certain sort of brain activity. The identity theory goes further to hold that mental states are *identical* to the associated brain states: these states are one and the same. This identification, unlike the behavioral thesis, is not grounded in an analysis of our concepts. Rather it is supposed to be an empirical claim, analogous to the claim that lightning is identical to electrical discharge, or that water

is identical to H_2O. In this way, the identity theory can be seen as driven by scientific developments, especially in neuroscience.

The classic statement of the identity theory was given by J. J. C. Smart (chapter 11) in 1959, refining and elaborating an earlier statement three years earlier by his colleague U. T. Place. Smart recognizes the strong intuitive resistance to the claim that mental states are brain states, especially in the case of conscious experiences, but he tries to defuse this resistance by addressing a number of objections. He suggests that mental concepts can be analyzed in a "topic-neutral way," so that nothing in these concepts alone dictates whether or not mental states are physical. This suggests that while it may not seem antecedently that mental states are physical, we can discover their physical nature through empirical science.

An objection to the identity theory, developed by Putnam (chapter 12), is that states such as pain cannot be identical to any particular brain states, since a creature such as a Martian might have pain without having the brain state in question. As Putnam put it, it is plausible that mental states are *multiply realizable*. If this is right, one cannot identify a mental state *type* such as being in pain with a physical state type, such as a specific sort of brain state. This still leaves open the possibility that one can identify mental state *tokens*, such as a specific pain of a subject, with physical state tokens, such as a specific biological state of that subject. Many philosophers think that the type identity theory is refuted by this objection, but that the token identity theory is left open.

The problems that Putnam raised for the identity theory and for behaviorism led him to embrace functionalism, which can be seen as a descendant of both. Broadly speaking, functionalism holds that mental states correspond to *functional* states: states of playing a certain role within the cognitive system. Putnam proposes a hypothesis that has come to be known as *machine functionalism*, according to which mental states are functional states of a computational machine. This makes mental states more abstract than any particular biological state and so allows the possibility of multiple realizations. It also allows a loose tie between mental states and behavior, without an absolute tie. This analogy between minds and machines has been very influential in contemporary philosophy of mind and cognitive science, and it has been developed in many different directions.

David Armstrong (chapter 13) puts forward a different form of functionalism, cast not in terms of machines but in terms of the general idea that mental states are defined in terms of their causal role. Specifically, he holds that the concept of a mental state is the concept of a state that is apt to be the cause of certain effects or apt to be the effect of certain causes. Where Putnam viewed his thesis as a sort of empirical hypothesis (like the identity theory), Armstrong puts forward his thesis as a sort of conceptual analysis (like logical behaviorism): it is a view about what we mean when we talk about the mind. Because of this, the view is often known as *analytic functionalism*. Unlike Putnam, Armstrong sees his view as supporting the identity theory rather than competing with it: if it turns out that in humans a specific brain state plays the causal role associated with pain, then that brain state is itself a pain.

This sort of functionalism is developed further by David Lewis (chapter 14). Lewis addresses two puzzles for the functionalist: Couldn't there be someone (a "madman") whose pain doesn't play the usual causal role for pain, and couldn't there be someone (a "Martian") whose pain has a physical realization quite different from ours? Lewis argues that analytic functionalism can handle both of these puzzles by distinguishing the *role* associated with pain from the physical property that *realizes* that role. In a given organism, pain should be identified with the physical property that realizes the role. In effect, pain is identified with different states in different organisms. In this way Lewis holds that his view can have the virtues of both the identity theory and functionalism.

That functionalism cannot deal with the "qualitative" aspects of conscious experience, such as the experience of seeing red or feeling pain, has often been objected to. Ned Block (chapter 15) develops the "absent qualia" objection, according to which a system could have the same functional states as a conscious system while having no

qualitative states at all. He argues for this thesis using a thought experiment involving a vast number of people causally organized to realize the given organization.

Martine Nida-Rümelin (chapter 16) develops a version of the "inverted qualia" objection, according to which two systems could have the same relevant functional states as a conscious system while having *different* qualitative states. If these objections are correct, then qualitative states are not identical to functional states, so functionalism is false.

FURTHER READING

Herbert Feigl's long and interesting 1958 article on the identity theory is reprinted with an afterword as Feigl 1967. The identity theory provoked a great deal of critical discussion in the 1960s, some of which is collected in Borst 1970; another collection is in Presley 1967. Influential objections to the identity theory are developed by Putnam (chapter 13) and Kripke (chapter 32). (chapter 33; see also Hill 1991) advocates a version of the type-identity theory.

There is an enormous and scattered literature on Balog functionalism; some relevant papers are collected in Biro and Shahan 1975. Putnam's version of functionalism is developed in a series of papers collected in Putnam 1975 and is repudiated in Putnam 1987. Armstrong's functionalism is developed at length in Armstrong 1968. The idea that mental states are defined in terms of a theory is also present in Sellars (chapter 42) and is developed further by Churchland (chapter 45). Searle (chapter 76) gives an argument against machine functionalism that is closely related to the absent qualia argument. Dennett (chapter 25) tries to deflate the idea of absent and inverted qualia, while Shoemaker 1975 and White 1986 give important defenses of functionalism against absent qualia objections. See also a number of related chapters on this topic (e.g. those by Nagel, Chalmers, and Kripke) in part 2 of this book. Shoemaker 1982 and Palmer 1999 discuss inverted qualia from a philosophical and empirical standpoint, respectively.

Armstrong, D. M. *A Materialist Theory of the Mind* (Routledge & Kegan Paul, 1968).

Biro, J. I., and Shahan, R. W., eds., *Mind, Brain and Function* (Oklahoma University Press, 1982).

Borst, C. V., ed., *The Mind/Brain Identity Theory* (Macmillan, 1970).

Feigl, H. *The "Mental" and the "Physical"* (University of Minnesota Press, 1967).

Hill, C. S. *Sensations: A Defense of Type Materialism* (Cambridge University Press, 1991).

Presley, C. P., ed., *The Identity Theory of Mind* (University of Queensland Press, 1967).

Palmer, S., "Color, consciousness, and the isomorphism constraint," *Behavioral and Brain Sciences* 22 (1999): pp. 1–21.

Putnam, H. *Mind, Language, and Reality* (Cambridge University Press, 1975).

_____ *Representation and Reality* (MIT Press, 1987).

Shoemaker, S., "Functionalism and qualia," *Philosophical Studies* 27 (1975): pp. 291–315. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

_____"The inverted spectrum," *Journal of Philosophy* 79 (1982): pp. 357–81. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

White, S., "Curse of the qualia," *Synthese* 68 (1986): pp. 333–68. Reprinted in *The Nature of Consciousness*, N. Block, O. Flanagan, and G. Güzeldere, eds., (Cambridge, MA: MIT Press, 1997).

D. Other Psychophysical Relations

A number of other views about the relationship between the mental and the physical have been put forward.

One important view is the *anomalous monism* of Donald Davidson (chapter 17). This view can be seen as an attempt to preserve materialism without any strong reduction

of the mental to the physical. In Davidson's view, any given mental event is identical to a physical event (a form of token identity theory), but there are no strict laws that connect mental events to physical events, and there are no strict laws governing mental events themselves. Davidson argues for this view by considering the distinct character of mental and physical concepts along with the causal connections between mental events and physical events.

There has been much discussion of whether the mental can be *reduced* to the physical, where this is understood as requiring more than the mere truth of materialism. Jerry Fodor (chapter 18) argues that in general, one cannot expect that the theories of a high-level "special science" should be reducible to the theories of a low-level science such as physics. Because of the many ways in which a high-level kind can be realized at a low level, the general principles in a high-level science cannot be captured by a low-level science except in a very complex and arbitrary way. This applies especially to the science of psychology, suggesting that one cannot expect that psychology can be reducible to physics, or even to neuroscience. Instead, it will always have a degree of autonomy.

Frank Jackson (chapter 19) argues for an important role for *conceptual analysis* in understanding the mental-physical relation. Like Davidson, Jackson holds that the mental supervenes on the physical: that is, any two possible systems with the same physical states will have the same mental states. Unlike Davidson, Jackson argues that this requires that mental concepts be analyzed in such a way so that there is a sort of a priori entailment from physical truths to mental truths. If this sort of conceptual analysis is possible, it becomes relatively easy to see how mental states can be physical. At the same time, the requirement of conceptual analyzability, if accepted, imposes a significant burden on the materialist.

One of the central problems in the metaphysics of the mind is the problem of mental causation: How can the mind affect the physical world? We have already seen that this is a major problem for dualism. More recently, it has been argued that mental causation also poses a problem for many forms of materialism. Jaegwon Kim (chapter 20) summarizes a number of different problems of mental causation. The first is a problem specifically for Davidson's anomalous monism: if there are no strict laws connecting mental states to physical states, how can mental states be causally relevant? The second is a problem specifically for externalist views about the mind (see section III. C): if mental states depend on factors outside the head, how can they affect behavior? The third is a problem raised for any view: given that one can give a full causal explanation of behavior in physical terms, how can mental states be causally relevant? This third problem can be called the "exclusion problem" since it suggests that mental states are excluded from causal explanation.

Another important view is *emergentism*, in which the mind is held to *emerge* from physical processes without being reducible to those processes. In recent philosophy, emergentism was popularized by the British emergentists of the early 20th century and especially C. D. Broad in *The Mind and its Place in Nature*. Jonardon Ganeri (chapter 21) argues that emergentism has roots in ancient Indian philosophy, and especially in the ideas of Bṛhaspati in the Cārvāka tradition of Hindu philosophy (around 600 BCE). Bṛhaspati holds that the physical elements earth, fire, air, and water are what's real, and that consciousness emerges from the combination of these elements. Ganeri argues that this view has interesting parallels with British emergentism and suggests an interestingly different conception of emergence.

Many discussions of the mind-body problem proceed by probing the concept of the mental, while taking the concept of the physical for granted. Barbara Montero (chapter 22) argues that the concept of the physical is unclear and problematic. As a result, the doctrine of physicalism—that the mind is fundamentally physical—has no clear content. Montero argues that we should move from physicalism to *post-physicalism*, where the key thesis is not that the mental is fundamentally physical but instead that the mental

is fundamentally nonmental. The opposing view is not dualism but the view that mentality is a fundamental ingredient of the universe. Montero argues that post-physicalism yields a clearer and more tractable debate.

FURTHER READING

C. D. Broad 1925 gives a classic statement of an emergentist view. The general program of British emergentists such as Broad is carefully described and analyzed by McLaughlin 1992. Beckermann, Flohr, and Kim 1992 is an excellent collection of papers on both emergence and reduction. Davidson's anomalous monism is analyzed in more depth by papers in McLaughlin and Lepore 1985 and Heil and Mele 1993. Fodor 1997 responds to Kim's arguments (as do other articles in the same volume). Bickle 1997 argues at length for a reductionist view. Wilson 1999 responds to Horgan's arguments. Important papers on supervenience are collected in Kim 1993. Jackson 1998 gives a more extensive treatment of the issues addressed in his paper here. Block and Stalnaker 1998 argue against Jackson's requirement of conceptual analyzability, and Chalmers and Jackson 2001 respond. Many aspects of the problem of mental causation are discussed by the papers in Heil and Mele 1993.

Beckermann, A., Flohr, H., and Kim, J., eds. *Emergence or Reduction? Prospects for Nonreductive Physicalism* (De Gruyter, 1992).

Bickle, J. *Psychoneural Reductionism: The New Wave* (Cambridge, MA: MIT Press, 1997).

Block, N., and Stalnaker, R., "Conceptual analysis, dualism, and the explanatory gap," *Philosophical Review* 108 (1999): pp. 1–46.

Broad, C. D. *The Mind and its Place in Nature* (Routledge & Kegan Paul, 1925).

Chalmers, D. J., and Jackson, F., "Conceptual analysis and reductive explanation," *Philosophical Review* 110 (2001): pp. 315–61.

Fodor, J., "Special sciences: Still autonomous after all these years," *Philosophical Perspectives* 11 (1997): pp. 149–63.

Heil, J., and Mele, A. *Mental Causation* (Oxford University Press, 1993).

Jackson, F. *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Oxford University Press, 1998).

Kim, J. *Supervenience and Mind* (Cambridge University Press, 1993).

McLaughlin, B. P., and Lepore, E. *Actions and Events* (Oxford: Blackwell Press, 1985).

Wilson, J., "How Superduper Does a Physicalist Supervenience Need to Be?" *Philosophical Quarterly* 49 (1999): pp. 33–52.

Meditations on First Philosophy

René Descartes

Second Meditation

The Nature of the Human Mind, and How It Is Better Known Than the Body

So serious are the doubts into which I have been thrown as a result of yesterday's meditation that I can neither put them out of my mind nor see any way of resolving them. It feels as if I have fallen unexpectedly into a deep whirlpool which tumbles me around so that I can neither stand on the bottom nor swim up to the top. Nevertheless I will make an effort and once more attempt the same path which I started on yesterday. Anything which admits of the slightest doubt I will set aside just as if I had found it to be wholly false; and I will proceed in this way until I recognize something certain, or, if nothing else, until I at least recognize for certain that there is no certainty. Archimedes used to demand just one firm and immovable point in order to shift the entire earth; so I too can hope for great things if I manage to find just one thing, however slight, that is certain and unshakeable. I will suppose then, that everything I see is spurious. I will believe that my memory tells me lies, and that none of the things that it reports ever happened. I have no senses. Body, shape, extension, movement and place are chimeras. So what remains true? Perhaps just the one fact that nothing is certain.

Yet apart from everything I have just listed, how do I know that there is not something else which does not allow even the slightest occasion for doubt? Is there not a God, or whatever I may call him, who puts into me[1] the thoughts I am now having? But why do I think this, since I myself may perhaps be the author of these thoughts? In that case am not I, at least, something? But I have just said that I have no senses and no body. This is the sticking point: what follows from this? Am I not so bound up with a body and with senses that I cannot exist without them? But I have convinced myself that there is absolutely nothing in the world, no sky, no earth, no minds, no bodies. Does it now follow that I too do not exist? No: if I convinced myself of something[2] then I certainly existed. But there is a deceiver of supreme power and cunning who is deliberately and constantly deceiving me. In that case I too undoubtedly exist, if he is deceiving me; and let him deceive me as much as he can, he will never bring it about that I am nothing so long as I think that I am something. So after considering everything very thoroughly, I must finally conclude that this proposition, *I am, I exist*, is necessarily true whenever it is put forward by me or conceived in my mind.

But I do not yet have a sufficient understanding of what this 'I' is, that now necessarily exists. So I must be on my guard against carelessly taking something else to be this 'I,' and so making a mistake in the very item of knowledge that I maintain is the most certain and evident of all. I will therefore go back and meditate on what I originally believed myself to be, before I embarked on this present train of thought. I will then subtract anything capable of being weakened, even minimally, by the arguments now introduced, so that what is left at the end may be exactly and only what is certain and unshakeable.

What then did I formerly think I was? A man. But what is a man? Shall I say 'a rational animal'? No; for then I should have to inquire what an animal is, what rationality is, and in this way one question would lead me down the slope to other harder ones, and I do not now have the time to waste on subtleties of this kind.

From Descartes, *Meditations on First Philosophy*, 1641. Translated in J. Cottingham, ed./trans., *Meditations on First Philosophy* (Cambridge University Press, 1985).

Instead I propose to concentrate on what came into my thoughts spontaneously and quite naturally whenever I used to consider what I was. Well, the first thought to come to mind was that I had a face, hands, arms and the whole mechanical structure of limbs which can be seen in a corpse, and which I called the body. The next thought was that I was nourished, that I moved about, and that I engaged in sense-perception and thinking; and these actions I attributed to the soul. But as to the nature of this soul, either I did not think about this or else I imagined it to be something tenuous, like a wind or fire or ether, which permeated my more solid parts. As to the body, however, I had no doubts about it, but thought I knew its nature distinctly. If I had tried to describe the mental conception I had of it, I would have expressed it as follows: by a body I understand whatever has a determinable shape and a definable location and can occupy a space in such a way as to exclude any other body; it can be perceived by touch, sight, hearing, taste or smell, and can be moved in various ways, not by itself but by whatever else comes into contact with it. For, according to my judgement, the power of self-movement, like the power of sensation or of thought, was quite foreign to the nature of a body; indeed, it was a source of wonder to me that certain bodies were found to contain faculties of this kind.

But what shall I now say that I am, when I am supposing that there is some supremely powerful and, if it is permissible to say so, malicious deceiver, who is deliberately trying to trick me in every way he can? Can I now assert that I possess even the most insignificant of all the attributes which I have just said belong to the nature of a body? I scrutinize them, think about them, go over them again, but nothing suggests itself; it is tiresome and pointless to go through the list once more. But what about the attributes I assigned to the soul? Nutrition or movement? Since now I do not have a body, these are mere fabrications. Sense-perception? This surely does not occur without a body, and besides, when asleep I have appeared to perceive through the senses many things which I afterwards realized I did not perceive through the senses at all. Thinking? At last I have discovered it—thought; this alone is inseparable from me. I am, I exist—that is certain. But for how long? For as long as I am thinking. For it could be that were I totally to cease from thinking, I should totally cease to exist. At present I am not admitting anything except what is necessarily true. I am, then, in the strict sense only a thing that thinks;[3] that is, I am a mind, or intelligence, or intellect, or reason—words whose meaning I have been ignorant of until now. But for all that I am a thing which is real and which truly exists. But what kind of a thing? As I have just said—a thinking thing.

What else am I? I will use my imagination.[4] I am not that structure of limbs which is called a human body. I am not even some thin vapour which permeates the limbs—a wind, fire, air, breath, or whatever I depict in my imagination; for these are things which I have supposed to be nothing. Let this supposition stand;[5] for all that I am still something. And yet may it not perhaps be the case that these very things which I am supposing to be nothing, because they are unknown to me, are in reality identical with the 'I' of which I am aware? I do not know, and for the moment I shall not argue the point, since I can make judgements only about things which are known to me. I know that I exist; the question is, what is this 'I' that I know? If the 'I' is understood strictly as we have been taking it, then it is quite certain that knowledge of it does not depend on things of whose existence I am as yet unaware; so it cannot depend on any of the things which I invent in my imagination. And this very word 'invent' shows me my mistake. It would indeed be a case of fictitious invention if I used my imagination to establish that I was something or other; for imagining is simply contemplating the shape or image of a corporeal thing. Yet now I know for certain both that I exist and at the same time that all such images and, in general, everything relating to the nature of body, could be mere dreams <and chimeras>. Once this point has been grasped, to say 'I will use my imagination to get to know more distinctly what I am' would seem to be as silly as saying 'I am now awake, and see some truth; but since my vision is not yet clear enough, I will deliberately fall asleep so that my dreams may provide a truer and clearer representation.' I thus realize that none of the things that the imagination enables me to grasp is at all relevant to this knowledge of myself which I possess, and that the mind must therefore be most carefully diverted from such things[6] if it is to perceive its own nature as distinctly as possible.

But what then am I? A thing that thinks. What is that? A thing that doubts, understands, affirms, denies, is willing, is unwilling, and also imagines and has sensory perceptions.

This is a considerable list, if everything on it belongs to me. But does it? Is it not one and the same 'I' who is now doubting almost

everything, who nonetheless understands some things, who affirms that this one thing is true, denies everything else, desires to know more, is unwilling to be deceived, imagines many things even involuntarily, and is aware of many things which apparently come from the senses? Are not all these things just as true as the fact that I exist, even if I am asleep all the time, and even if he who created me is doing all he can to deceive me? Which of all these activities is distinct from my thinking? Which of them can be said to be separate from myself? The fact that it is I who am doubting and understanding and willing is so evident that I see no way of making it any clearer. But it is also the case that the 'I' who imagines is the same 'I.' For even if, as I have supposed, none of the objects of imagination are real, the power of imagination is something which really exists and is part of my thinking. Lastly, it is also the same 'I' who has sensory perceptions, or is aware of bodily things as it were through the senses. For example, I am now seeing light, hearing a noise, feeling heat. But I am asleep, so all this is false. Yet I certainly *seem* to see, to hear, and to be warmed. This cannot be false; what is called 'having a sensory perception' is strictly just this, and in this restricted sense of the term it is simply thinking.

From all this I am beginning to have a rather better understanding of what I am. But it still appears—and I cannot stop thinking this—that the corporeal things of which images are formed in my thought, and which the senses investigate, are known with much more distinctness than this puzzling 'I' which cannot be pictured in the imagination. And yet it is surely surprising that I should have a more distinct grasp of things which I realize are doubtful, unknown and foreign to me, than I have of that which is true and known—my own self. But I see what it is: my mind enjoys wandering off and will not yet submit to being restrained within the bounds of truth. Very well then; just this once let us give it a completely free rein, so that after a while, when it is time to tighten the reins, it may more readily submit to being curbed.

Let us consider the things which people commonly think they understand most distinctly of all; that is, the bodies which we touch and see. I do not mean bodies in general—for general perceptions are apt to be somewhat more confused—but one particular body. Let us take, for example, this piece of wax. It has just been taken from the honeycomb; it has not yet quite lost the taste of the honey; it retains some of the scent of the flowers from which it was gathered; its colour, shape and size are plain to see; it is hard, cold and can be handled without difficulty; if you rap it with your knuckle it makes a sound. In short, it has everything which appears necessary to enable a body to be known as distinctly as possible. But even as I speak, I put the wax by the fire, and look: the residual taste is eliminated, the smell goes away, the colour changes, the shape is lost, the size increases; it becomes liquid and hot; you can hardly touch it, and if you strike it, it no longer makes a sound. But does the same wax remain? It must be admitted that it does; no one denies it, no one thinks otherwise. So what was it in the wax that I understood with such distinctness? Evidently none of the features which I arrived at by means of the senses; for whatever came under taste, smell, sight, touch or hearing has now altered—yet the wax remains.

Perhaps the answer lies in the thought which now comes to my mind; namely, the wax was not after all the sweetness of the honey, or the fragrance of the flowers, or the whiteness, or the shape, or the sound, but was rather a body which presented itself to me in these various forms a little while ago, but which now exhibits different ones. But what exactly is it that I am now imagining? Let us concentrate, take away everything which does not belong to the wax, and see what is left: merely something extended, flexible and changeable. But what is meant here by 'flexible' and 'changeable'? Is it what I picture in my imagination: that this piece of wax is capable of changing from a round shape to a square shape, or from a square shape to a triangular shape? Not at all; for I can grasp that the wax is capable of countless changes of this kind, yet I am unable to run through this immeasurable number of changes in my imagination, from which it follows that it is not the faculty of imagination that gives me my grasp of the wax as flexible and changeable. And what is meant by 'extended'? Is the extension of the wax also unknown? For it increases if the wax melts, increases again if it boils, and is greater still if the heat is increased. I would not be making a correct judgement about the nature of wax unless I believed it capable of being extended in many more different ways than I will ever encompass in my imagination. I must therefore admit that the nature of this piece of wax is in no way revealed by my imagination, but is perceived by the mind alone. (I am speaking of this particular piece of wax; the point is even clearer with regard to wax in general.) But what is this wax which is perceived by the mind alone?[7] It is of course the same wax which I see,

which I touch, which I picture in my imagination, in short the same wax which I thought it to be from the start. And yet, and here is the point, the perception I have of it[8] is a case not of vision or touch or imagination—nor has it ever been, despite previous appearances—but of purely mental scrutiny; and this can be imperfect and confused, as it was before, or clear and distinct as it is now, depending on how carefully I concentrate on what the wax consists in.

But as I reach this conclusion I am amazed at how <weak and> prone to error my mind is. For although I am thinking about these matters within myself, silently and without speaking, nonetheless the actual words bring me up short, and I am almost tricked by ordinary ways of talking. We say that we see the wax itself, if it is there before us, not that we judge it to be there from its colour or shape; and this might lead me to conclude without more ado that knowledge of the wax comes from what the eye sees, and not from the scrutiny of the mind alone. But then if I look out of the window and see men crossing the square, as I just happen to have done, I normally say that I see the men themselves, just as I say that I see the wax. Yet do I see any more than hats and coats which could conceal automatons? I *judge* that they are men. And so something which I thought I was seeing with my eyes is in fact grasped solely by the faculty of judgement which is in my mind.

However, one who wants to achieve knowledge above the ordinary level should feel ashamed at having taken ordinary ways of talking as a basis for doubt. So let us proceed, and consider on which occasion my perception of the nature of the wax was more perfect and evident. Was it when I first looked at it, and believed I knew it by my external senses, or at least by what they call the 'common' sense[9]—that is, the power of imagination? Or is my knowledge more perfect now, after a more careful investigation of the nature of the wax and of the means by which it is known? Any doubt on this issue would clearly be foolish; for what distinctness was there in my earlier perception? Was there anything in it which an animal could not possess? But when I distinguish the wax from its outward forms—take the clothes off, as it were, and consider it naked—then although my judgement may still contain errors, at least my perception now requires a human mind.

But what am I to say about this mind, or about myself? (So far, remember, I am not admitting that there is anything else in me except a mind.) What, I ask, is this 'I' which seems to perceive the wax so distinctly? Surely my awareness of my own self is not merely much truer and more certain than my awareness of the wax, but also much more distinct and evident. For if I judge that the wax exists from the fact that I see it, clearly this same fact entails much more evidently that I myself also exist. It is possible that what I see is not really the wax; it is possible that I do not even have eyes with which to see anything. But when I see, or think I see (I am not here distinguishing the two), it is simply not possible that I who am now thinking am not something. By the same token, if I judge that the wax exists from the fact that I touch it, the same result follows, namely that I exist. If I judge that it exists from the fact that I imagine it, or for any other reason, exactly the same thing follows. And the result that I have grasped in the case of the wax may be applied to everything else located outside me. Moreover, if my perception of the wax seemed more distinct[10] after it was established not just by sight or touch but by many other considerations, it must be admitted that I now know myself even more distinctly. This is because every consideration whatsoever which contributes to my perception of the wax, or of any other body, cannot but establish even more effectively the nature of my own mind. But besides this, there is so much else in the mind itself which can serve to make my knowledge of it more distinct, that it scarcely seems worth going through the contributions made by considering bodily things.

I see that without any effort I have now finally got back to where I wanted. I now know that even bodies are not strictly perceived by the senses or the faculty of imagination but by the intellect alone, and that this perception derives not from their being touched or seen but from their being understood; and in view of this I know plainly that I can achieve an easier and more evident perception of my own mind than of anything else. But since the habit of holding on to old opinions cannot be set aside so quickly, I should like to stop here and meditate for some time on this new knowledge I have gained, so as to fix it more deeply in my memory.

Sixth Meditation

The Existence of Material Things, and the Real Distinction between Mind and Body[11]

It remains for me to examine whether material things exist. And at least I now know they are capable of existing, in so far as they are the

subject-matter of pure mathematics, since I perceive them clearly and distinctly. For there is no doubt that God is capable of creating everything that I am capable of perceiving in this manner; and I have never judged that something could not be made by him except on the grounds that there would be a contradiction in my perceiving it distinctly. The conclusion that material things exist is also suggested by the faculty of imagination, which I am aware of using when I turn my mind to material things. For when I give more attentive consideration to what imagination is, it seems to be nothing else but an application of the cognitive faculty to a body which is intimately present to it, and which therefore exists.

To make this clear, I will first examine the difference between imagination and pure understanding. When I imagine a triangle, for example, I do not merely understand that it is a figure bounded by three lines, but at the same time I also see the three lines with my mind's eye as if they were present before me; and this is what I call imagining. But if I want to think of a chiliagon, although I understand that it is a figure consisting of a thousand sides just as well as I understand the triangle to be a three-sided figure, I do not in the same way imagine the thousand sides or see them as if they were present before me. It is true that since I am in the habit of imagining something whenever I think of a corporeal thing, I may construct in my mind a confused representation of some figure; but it is clear that this is not a chiliagon. For it differs in no way from the representation I should form if I were thinking of a myriagon, or any figure with very many sides. Moreover, such a representation is useless for recognizing the properties which distinguish a chiliagon from other polygons. But suppose I am dealing with a pentagon: I can of course understand the figure of a pentagon, just as I can the figure of a chiliagon, without the help of the imagination; but I can also imagine a pentagon, by applying my mind's eye to its five sides and the area contained within them. And in doing this I notice quite clearly that imagination requires a peculiar effort of mind which is not required for understanding; this additional effort of mind clearly shows the difference between imagination and pure understanding.

Besides this, I consider that this power of imagining which is in me, differing as it does from the power of understanding, is not a necessary constituent of my own essence, that is, of the essence of my mind. For if I lacked it, I should undoubtedly remain the same individual as I now am; from which it seems to follow that it depends on something distinct from myself. And I can easily understand that, if there does exist some body to which the mind is so joined that it can apply itself to contemplate it, as it were, whenever it pleases, then it may possibly be this very body that enables me to imagine corporeal things. So the difference between this mode of thinking and pure understanding may simply be this: when the mind understands, it in some way turns towards itself and inspects one of the ideas which are within it; but when it imagines, it turns towards the body and looks at something in the body which conforms to an idea understood by the mind or perceived by the senses. I can, as I say, easily understand that this is how imagination comes about, if the body exists; and since there is no other equally suitable way of explaining imagination that comes to mind, I can make a probable conjecture that the body exists. But this is only a probability; and despite a careful and comprehensive investigation, I do not yet see how the distinct idea of corporeal nature which I find in my imagination can provide any basis for a necessary inference that some body exists.

But besides that corporeal nature which is the subject-matter of pure mathematics, there is much else that I habitually imagine, such as colours, sounds, tastes, pain and so on—though not so distinctly. Now I perceive these things much better by means of the senses, which is how, with the assistance of memory, they appear to have reached the imagination. So in order to deal with them more fully, I must pay equal attention to the senses, and see whether the things which are perceived by means of that mode of thinking which I call 'sensory perception' provide me with any sure argument for the existence of corporeal things.

To begin with, I will go back over all the things which I previously took to be perceived by the senses, and reckoned to be true; and I will go over my reasons for thinking this. Next, I will set out my reasons for subsequently calling these things into doubt. And finally I will consider what I should now believe about them.

First of all then, I perceived by my senses that I had a head, hands, feet and other limbs making up the body which I regarded as part of myself, or perhaps even as my whole self. I also perceived by my senses that this body was situated among many other bodies which could affect it in various favourable or unfavourable ways; and I gauged the favourable effects by

a sensation of pleasure, and the unfavourable ones by a sensation of pain. In addition to pain and pleasure, I also had sensations within me of hunger, thirst, and other such appetites, and also of physical propensities towards cheerfulness, sadness, anger and similar emotions. And outside me, besides the extension, shapes and movements of bodies, I also had sensations of their hardness and heat, and of the other tactile qualities. In addition, I had sensations of light, colours, smells, tastes and sounds, the variety of which enabled me to distinguish the sky, the earth, the seas, and all other bodies, one from another. Considering the ideas of all these qualities which presented themselves to my thought, although the ideas were, strictly speaking, the only immediate objects of my sensory awareness, it was not unreasonable for me to think that the items which I was perceiving through the senses were things quite distinct from my thought, namely bodies which produced the ideas. For my experience was that these ideas came to me quite without my consent, so that I could not have sensory awareness of any object, even if I wanted to, unless it was present to my sense organs; and I could not avoid having sensory awareness of it when it was present. And since the ideas perceived by the senses were much more lively and vivid and even, in their own way, more distinct than any of those which I deliberately formed through meditating or which I found impressed on my memory, it seemed impossible that they should have come from within me; so the only alternative was that they came from other things. Since the sole source of my knowledge of these things was the ideas themselves, the supposition that the things resembled the ideas was bound to occur to me. In addition, I remembered that the use of my senses had come first, while the use of my reason came only later; and I saw that the ideas which I formed myself were less vivid than those which I perceived with the senses and were, for the most part, made up of elements of sensory ideas. In this way I easily convinced myself that I had nothing at all in the intellect which I had not previously had in sensation. As for the body which by some special right I called 'mine my belief that this body, more than any other, belonged to me had some justification. For I could never be separated from it, as I could from other bodies; and I felt all my appetites and emotions in, and on account of, this body; and finally, I was aware of pain and pleasurable ticklings in parts of this body, but not in other bodies external to it. But why should that curious sensation of pain give rise to a particular distress of mind; or why should a certain kind of delight follow on a tickling sensation? Again, why should that curious tugging in the stomach which I call hunger tell me that I should eat, or a dryness of the throat tell me to drink, and so on? I was not able to give any explanation of all this, except that nature taught me so. For there is absolutely no connection (at least that I can understand) between the tugging sensation and the decision to take food, or between the sensation of something causing pain and the mental apprehension of distress that arises from that sensation. These and other judgements that I made concerning sensory objects, I was apparently taught to make by nature; for I had already made up my mind that this was how things were, before working out any arguments to prove it.

Later on, however, I had many experiences which gradually undermined all the faith I had had in the senses. Sometimes towers which had looked round from a distance appeared square from close up; and enormous statues standing on their pediments did not seem large when observed from the ground. In these and countless other such cases, I found that the judgements of the external senses were mistaken. And this applied not just to the external senses but to the internal senses as well. For what can be more internal than pain? And yet I had heard that those who had had a leg or an arm amputated sometimes still seemed to feel pain intermittently in the missing part of the body. So even in my own case it was apparently not quite certain that a particular limb was hurting, even if I felt pain in it. To these reasons for doubting, I recently added two very general ones.[12] The first was that every sensory experience I have ever thought I was having while awake I can also think of myself as sometimes having while asleep; and since I do not believe that what I seem to perceive in sleep comes from things located outside me, I did not see why I should be any more inclined to believe this of what I think I perceive while awake. The second reason for doubt was that since I did not know the author of my being (or at least was pretending not to), I saw nothing to rule out the possibility that my natural constitution made me prone to error even in matters which seemed to me most true. As for the reasons for my previous confident belief in the truth of the things perceived by the senses, I had no trouble in refuting them. For since I apparently had natural impulses towards many things which reason told me to

avoid, I reckoned that a great deal of confidence should not be placed in what I was taught by nature. And despite the fact that the perceptions of the senses were not dependent on my will, I did not think that I should on that account infer that they proceeded from things distinct from myself, since I might perhaps have a faculty not yet known to me which produced them.[13]

But now, when I am beginning to achieve a better knowledge of myself and the author of my being, although I do not think I should heedlessly accept everything I seem to have acquired from the senses, neither do I think that everything should be called into doubt.

First, I know that everything which I clearly and distinctly understand is capable of being created by God so as to correspond exactly with my understanding of it. Hence the fact that I can clearly and distinctly understand one thing apart from another is enough to make me certain that the two things are distinct, since they are capable of being separated, at least by God. The question of what kind of power is required to bring about such a separation does not affect the judgement that the two things are distinct. Thus, simply by knowing that I exist and seeing at the same time that absolutely nothing else belongs to my nature or essence except that I am a thinking thing, I can infer correctly that my essence consists solely in the fact that I am a thinking thing. It is true that I may have (or, to anticipate, that I certainly have) a body that is very closely joined to me. But nevertheless, on the one hand I have a clear and distinct idea of myself, in so far as I am simply a thinking, non-extended thing; and on the other hand I have a distinct idea of body,[14] in so far as this is simply an extended, non-thinking thing. And accordingly, it is certain that I[15] am really distinct from my body, and can exist without it.

Besides this, I find in myself faculties for certain special modes of thinking,[16] namely imagination and sensory perception. Now I can clearly and distinctly understand myself as a whole without these faculties; but I cannot, conversely, understand these faculties without me, that is, without an intellectual substance to inhere in. This is because there is an intellectual act included in their essential definition; and hence I perceive that the distinction between them and myself corresponds to the distinction between the modes of a thing and the thing itself.[17] Of course I also recognize that there are other faculties (like those of changing position, of taking on various shapes, and so on) which, like sensory perception and imagination, cannot be understood apart from some substance for them to inhere in, and hence cannot exist without it. But it is clear that these other faculties, if they exist, must be in a corporeal or extended substance and not an intellectual one; for the clear and distinct conception of them includes extension, but does not include any intellectual act whatsoever. Now there is in me a passive faculty of sensory perception, that is, a faculty for receiving and recognizing the ideas of sensible objects; but I could not make use of it unless there was also an active faculty, either in me or in something else, which produced or brought about these ideas. But this faculty cannot be in me, since clearly it presupposes no intellectual act on my part,[18] and the ideas in question are produced without my cooperation and often even against my will. So the only alternative is that it is in another substance distinct from me—a substance which contains either formally or eminently all the reality which exists objectively[19] in the ideas produced by this faculty (as I have just noted). This substance is either a body, that is, a corporeal nature, in which case it will contain formally <and in fact> everything which is to be found objectively <or representatively> in the ideas; or else it is God, or some creature more noble than a body, in which case it will contain eminently whatever is to be found in the ideas. But since God is not a deceiver, it is quite clear that he does not transmit the ideas to me either directly from himself, or indirectly, via some creature which contains the objective reality of the ideas not formally but only eminently. For God has given me no faculty at all for recognizing any such source for these ideas; on the contrary, he has given me a great propensity to believe that they are produced by corporeal things. So I do not see how God could be understood to be anything but a deceiver if the ideas were transmitted from a source other than corporeal things. It follows that corporeal things exist. They may not all exist in a way that exactly corresponds with my sensory grasp of them, for in many cases the grasp of the senses is very obscure and confused. But at least they possess all the properties which I clearly and distinctly understand, that is, all those which, viewed in general terms, are comprised within the subject-matter of pure mathematics.

What of the other aspects of corporeal things which are either particular (for example that the sun is of such and such a size or shape), or less clearly understood, such as light or sound or pain, and so on? Despite the high degree of doubt and uncertainty involved here, the very

fact that God is not a deceiver, and the consequent impossibility of there being any falsity in my opinions which cannot be corrected by some other faculty supplied by God, offers me a sure hope that I can attain the truth even in these matters. Indeed, there is no doubt that everything that I am taught by nature contains some truth. For if nature is considered in its general aspect, then I understand by the term nothing other than God himself, or the ordered system of created things established by God. And by my own nature in particular I understand nothing other than the totality of things bestowed on me by God.

There is nothing that my own nature teaches me more vividly than that I have a body, and that when I feel pain there is something wrong with the body, and that when I am hungry or thirsty the body needs food and drink, and so on. So I should not doubt that there is some truth in this.

Nature also teaches me, by these sensations of pain, hunger, thirst and so on, that I am not merely present in my body as a sailor is present in a ship,[19] but that I am very closely joined and, as it were, intermingled with it, so that I and the body form a unit. If this were not so, I, who am nothing but a thinking thing, would not feel pain when the body was hurt, but would perceive the damage purely by the intellect, just as a sailor perceives by sight if anything in his ship is broken. Similarly, when the body needed food or drink, I should have an explicit understanding of the fact, instead of having confused sensations of hunger and thirst. For these sensations of hunger, thirst, pain and so on are nothing but confused modes of thinking which arise from the union and, as it were, intermingling of the mind with the body.

I am also taught by nature that various other bodies exist in the vicinity of my body, and that some of these are to be sought out and others avoided. And from the fact that I perceive by my senses a great variety of colours, sounds, smells and tastes, as well as differences in heat, hardness and the like, I am correct in inferring that the bodies which are the source of these various sensory perceptions possess differences corresponding to them, though perhaps not resembling them. Also, the fact that some of the perceptions are agreeable to me while others are disagreeable makes it quite certain that my body, or rather my whole self, in so far as I am a combination of body and mind, can be affected by the various beneficial or harmful bodies which surround it.

There are, however, many other things which I may appear to have been taught by nature, but which in reality I acquired not from nature but from a habit of making ill-considered judgements; and it is therefore quite possible that these are false. Cases in point are the belief that any space in which nothing is occurring to stimulate my senses must be empty; or that the heat in a body is something exactly resembling the idea of heat which is in me; or that when a body is white or green, the selfsame whiteness or greenness which I perceive through my senses is present in the body; or that in a body which is bitter or sweet there is the selfsame taste which I experience, and so on; or, finally, that stars and towers and other distant bodies have the same size and shape which they present to my senses, and other examples of this kind. But to make sure that my perceptions in this matter are sufficiently distinct, I must more accurately define exactly what I mean when I say that I am taught something by nature. In this context I am taking nature to be something more limited than the totality of things bestowed on me by God. For this includes many things that belong to the mind alone—for example my perception that what is done cannot be undone, and all other things that are known by the natural light;[20] but at this stage I am not speaking of these matters. It also includes much that relates to the body alone, like the tendency to move in a downward direction, and so on; but I am not speaking of these matters either. My sole concern here is with what God has bestowed on me as a combination of mind and body. My nature, then, in this limited sense, does indeed teach me to avoid what induces a feeling of pain and to seek out what induces feelings of pleasure, and so on. But it does not appear to teach us to draw any conclusions from these sensory perceptions about things located outside us without waiting until the intellect has examined[21] the matter. For knowledge of the truth about such things seems to belong to the mind alone, not to the combination of mind and body. Hence, although a star has no greater effect on my eye than the flame of a small light, that does not mean that there is any real or positive inclination in me to believe that the star is no bigger than the light; I have simply made this judgement from childhood onwards without any rational basis. Similarly, although I feel heat when I go near a fire and feel pain when I go too near, there is no convincing argument for supposing that there is something in the fire which resembles the heat, any more than for supposing that there

is something which resembles the pain. There is simply reason to suppose that there is something in the fire, whatever it may eventually turn out to be, which produces in us the feelings of heat or pain. And likewise, even though there is nothing in any given space that stimulates the senses, it does not follow that there is no body there. In these cases and many others I see that I have been in the habit of misusing the order of nature. For the proper purpose of the sensory perceptions given me by nature is simply to inform the mind of what is beneficial or harmful for the composite of which the mind is a part; and to this extent they are sufficiently clear and distinct. But I misuse them by treating them as reliable touchstones for immediate judgements about the essential nature of the bodies located outside us; yet this is an area where they provide only very obscure information.

I have already looked in sufficient detail at how, notwithstanding the goodness of God, it may happen that my judgements are false. But a further problem now comes to mind regarding those very things which nature presents to me as objects which I should seek out or avoid, and also regarding the internal sensations, where I seem to have detected errors[22]—e.g. when someone is tricked by the pleasant taste of some food into eating the poison concealed inside it. Yet in this case, what the man's nature urges him to go for is simply what is responsible for the pleasant taste, and not the poison, which his nature knows nothing about. The only inference that can be drawn from this is that his nature is not omniscient. And this is not surprising, since man is a limited thing, and so it is only fitting that his perfection should be limited.

And yet it is not unusual for us to go wrong even in cases where nature does urge us towards something. Those who are ill, for example, may desire food or drink that will shortly afterwards turn out to be bad for them. Perhaps it may be said that they go wrong because their nature is disordered, but this does not remove the difficulty. A sick man is no less one of God's creatures than a healthy one, and it seems no less a contradiction to suppose that he has received from God a nature which deceives him. Yet a clock constructed with wheels and weights observes all the laws of its nature just as closely when it is badly made and tells the wrong time as when it completely fulfils the wishes of the clockmaker. In the same way, I might consider the body of a man as a kind of machine equipped with and made up of bones, nerves, muscles, veins, blood and skin in such a way that, even if there were no mind in it, it would still perform all the same movements as it now does in those cases where movement is not under the control of the will or, consequently, of the mind.[23] I can easily see that if such a body suffers from dropsy, for example, and is affected by the dryness of the throat which normally produces in the mind the sensation of thirst, the resulting condition of the nerves and other parts will dispose the body to take a drink, with the result that the disease will be aggravated. Yet this is just as natural as the body's being stimulated by a similar dryness of the throat to take a drink when there is no such illness and the drink is beneficial. Admittedly, when I consider the purpose of the clock, I may say that it is departing from its nature when it does not tell the right time; and similarly when I consider the mechanism of the human body, I may think that, in relation to the movements which normally occur in it, it too is deviating from its nature if the throat is dry at a time when drinking is not beneficial to its continued health. But I am well aware that 'nature' as I have just used it has a very different significance from 'nature' in the other sense. As I have just used it, 'nature' is simply a label which depends on my thought; it is quite extraneous to the things to which it is applied, and depends simply on my comparison between the idea of a sick man and a badly-made clock, and the idea of a healthy man and a well-made clock. But by 'nature' in the other sense I understand something which is really to be found in the things themselves; in this sense, therefore, the term contains something of the truth.

When we say, then, with respect to the body suffering from dropsy, that it has a disordered nature because it has a dry throat and yet does not need drink, the term 'nature' is here used merely as an extraneous label. However, with respect to the composite, that is, the mind united with this body, what is involved is not a mere label, but a true error of nature, namely that it is thirsty at a time when drink is going to cause it harm. It thus remains to inquire how it is that the goodness of God does not prevent nature, in this sense, from deceiving us.

The first observation I make at this point is that there is a great difference between the mind and the body, inasmuch as the body is by its very nature always divisible, while the mind is utterly indivisible. For when I consider the mind, or myself in so far as I am merely a thinking thing, I am unable to distinguish any parts within myself; I understand myself to be

something quite single and complete. Although the whole mind seems to be united to the whole body, I recognize that if a foot or arm or any other part of the body is cut off, nothing has thereby been taken away from the mind. As for the faculties of willing, of understanding, of sensory perception and so on, these cannot be termed parts of the mind, since it is one and the same mind that wills, and understands and has sensory perceptions. By contrast, there is no corporeal or extended thing that I can think of which in my thought I cannot easily divide into parts; and this very fact makes me understand that it is divisible. This one argument would be enough to show me that the mind is completely different from the body, even if I did not already know as much from other considerations.

My next observation is that the mind is not immediately affected by all parts of the body, but only by the brain, or perhaps just by one small part of the brain, namely the part which is said to contain the 'common' sense.[24] Every time this part of the brain is in a given state, it presents the same signals to the mind, even though the other parts of the body may be in a different condition at the time. This is established by countless observations, which there is no need to review here.

I observe, in addition, that the nature of the body is such that whenever any part of it is moved by another part which is some distance away, it can always be moved in the same fashion by any of the parts which lie in between, even if the more distant part does nothing. For example, in a cord ABCD, if one end D is pulled so that the other end A moves, the exact same movement could have been brought about if one of the intermediate points B or C had been pulled, and D had not moved at all. In similar fashion, when I feel a pain in my foot, physiology tells me that this happens by means of nerves distributed throughout the foot, and that these nerves are like cords which go from the foot right up to the brain. When the nerves are pulled in the foot, they in turn pull on inner parts of the brain to which they are attached, and produce a certain motion in them; and nature had laid it down that this motion should produce in the mind a sensation of pain, as occurring in the foot. But since these nerves, in passing from the foot to the brain, must pass through the calf, the thigh, the lumbar region, the back and the neck, it can happen that, even if it is not the part in the foot but one of the intermediate parts which is being pulled, the same motion will occur in the brain as occurs when the foot is hurt, and so it will necessarily come about that the mind feels the same sensation of pain. And we must suppose the same thing happens with regard to any other sensation.

My final observation is that any given movement occurring in the part of the brain that immediately affects the mind produces just one corresponding sensation; and hence the best system that could be devised is that it should produce the one sensation which, of all possible sensations, is most especially and most frequently conducive to the preservation of the healthy man. And experience shows that the sensations which nature has given us are all of this kind; and so there is absolutely nothing to be found in them that does not bear witness to the power and goodness of God. For example, when the nerves in the foot are set in motion in a violent and unusual manner, this motion, by way of the spinal cord, reaches the inner parts of the brain, and there gives the mind its signal for having a certain sensation, namely the sensation of a pain as occurring in the foot. This stimulates the mind to do its best to get rid of the cause of the pain, which it takes to be harmful to the foot. It is true that God could have made the nature of man such that this particular motion in the brain indicated something else to the mind; it might, for example, have made the mind aware of the actual motion occurring in the brain, or in the foot, or in any of the intermediate regions; or it might have indicated something else entirely. But there is nothing else which would have been so conducive to the continued well-being of the body. In the same way, when we need drink, there arises a certain dryness in the throat; this sets in motion the nerves of the throat, which in turn move the inner parts of the brain. This motion produces in the mind a sensation of thirst, because the most useful thing for us to know about the whole business is that we need drink in order to stay healthy. And so it is in the other cases.

It is quite clear from all this that, notwithstanding the immense goodness of God, the nature of man as a combination of mind and body is such that it is bound to mislead him from time to time. For there may be some occurrence, not in the foot but in one of the other areas through which the nerves travel in their route from the foot to the brain, or even in the brain itself; and if this cause produces the same motion which is generally produced by injury to the foot, then pain will be felt as if it were in the foot. This deception of the senses is natural, because a given motion in the brain must always produce

the same sensation in the mind; and the origin of the motion in question is much more often going to be something which is hurting the foot, rather than something existing elsewhere. So it is reasonable that this motion should always indicate to the mind a pain in the foot rather than in any other part of the body. Again, dryness of the throat may sometimes arise not, as it normally does, from the fact that a drink is necessary to the health of the body, but from some quite opposite cause, as happens in the case of the man with dropsy. Yet it is much better that it should mislead on this occasion than that it should always mislead when the body is in good health. And the same goes for the other cases.

This consideration is the greatest help to me, not only for noticing all the errors to which my nature is liable, but also for enabling me to correct or avoid them without difficulty. For I know that in matters regarding the well-being of the body, all my senses report the truth much more frequently than not. Also, I can almost always make use of more than one sense to investigate the same thing; and in addition, I can use both my memory, which connects present experiences with preceding ones, and my intellect, which has by now examined all the causes of error. Accordingly, I should not have any further fears about the falsity of what my senses tell me every day; on the contrary, the exaggerated doubts of the last few days should be dismissed as laughable. This applies especially to the principal reason for doubt, namely my inability to distinguish between being asleep and being awake. For I now notice that there is a vast difference between the two, in that dreams are never linked by memory with all the other actions of life as waking experiences are. If, while I am awake, anyone were suddenly to appear to me and then disappear immediately, as happens in sleep, so that I could not see where he had come from or where he had gone to, it would not be unreasonable for me to judge that he was a ghost, or a vision created in my brain,[25] rather than a real man. But when I distinctly see where things come from and where and when they come to me, and when I can connect my perceptions of them with the whole of the rest of my life without a break, then I am quite certain that when I encounter these things I am not asleep but awake. And I ought not to have even the slightest doubt of their reality if, after calling upon all the senses as well as my memory and my intellect in order to check them, I receive no conflicting reports from any of these sources. For from the fact that God is not a deceiver it follows that in cases like these I am completely free from error. But since the pressure of things to be done does not always allow us to stop and make such a meticulous check, it must be admitted that in this human life we are often liable to make mistakes about particular things, and we must acknowledge the weakness of our nature.

NOTES

1. '. . . puts into my mind' (French version).
2. '. . . or thought anything at all' (French version).
3. The word 'only' is most naturally taken as going with 'a thing that thinks and this interpretation is followed in the French version. When discussing this passage with Gassendi, however, Descartes suggests that he meant the 'only' to govern 'in the strict sense'; cf. AT IXA 215; CSM II 276.
4. '. . . to see if I am not something more' (added in French version).
5. Lat. *maneat* ('let it stand'), first edition. The second edition has the indicative *manet:* 'The proposition still stands, *viz.* that I am nonetheless something.' The French version reads: 'without changing this supposition, I find that I am still certain that I am something.'
6. '. . . from this manner of conceiving things' (French version).
7. '. . . which can be conceived only by the understanding or the mind' (French version).
8. '. . . or rather the act whereby it is perceived' (added in French version).
9. See note 24 below.
10. The French version has 'more clear and distinct' and, at the end of this sentence, 'more evidently, distinctly and clearly.'
11. '. . . between the soul and body of a man' (French version).
12. Cf. Med. 1.
13. Cf. Med. III.
14. The Latin term *corpus* as used here by Descartes is ambiguous as between 'body' (i.e., corporeal matter in general) and 'the body' (i.e., this particular body of mine). The French version preserves the ambiguity.
15. '. . . that is, my soul, by which I am what I am' (added in French version).
16. '. . . certain modes of thinking which are quite special and distinct from me' (French version).
17. '. . . between the shapes, movements and other modes or accidents of a body and the body which supports them' (French version).
18. '. . . cannot be in me in so far as I am merely a thinking thing, since it does not presuppose any thought on my part' (French version).
19. '. . . as a pilot in his ship' (French version).
20. '. . . without any help from the body' (added in French version).
21. '. . . carefully and maturely examined' (French version).

22. '. . . and thus seem to have been directly deceived by my nature' (added in French version).
23. '. . . but occurs merely as a result of the disposition of the organs' (French version).
24. The supposed faculty which integrates the data from the five specialized senses (the notion goes back ultimately to Aristotle). 'The seat of the common sense must be very mobile, to receive all the impressions coming from the senses, but must be moveable only by the spirits which transmit these impressions. Only the *conarion* [pineal gland] fits these conditions' (letter to Mersenne, 21 April 1641).
25. '. . . like those that are formed in the brain when I sleep' (added in French version).

The Passions of the Soul

René Descartes

17. The Functions of the Soul

Having thus considered all the functions belonging solely to the body, it is easy to recognize that there is nothing in us which we must attribute to our soul except our thoughts. These are of two principal kinds, some being actions of the soul and others its passions. Those I call its actions are all our volitions, for we experience them as proceeding directly from our soul and as seeming to depend on it alone. On the other hand, the various perceptions or modes of knowledge present in us may be called its passions, in a general sense, for it is often not our soul which makes them such as they are, and the soul always receives them from the things that are represented by them.

18. The Will

Our volitions, in turn, are of two sorts. One consists of the actions of the soul which terminate in the soul itself, as when we will to love God or, generally speaking, to apply our mind to some object which is not material. The other consists of actions which terminate in our body, as when our merely willing to walk has the consequence that our legs move and we walk.

19. Perception

Our perceptions are likewise of two sorts: some have the soul as their cause, others the body. Those having the soul as their cause are the perceptions of our volitions and of all the imaginings or other thoughts which depend on them. For it is certain that we cannot will anything without thereby perceiving that we are willing it. And although willing something is an action with respect to our soul, the perception of such willing may be said to be a passion in the soul. But because this perception is really one and the same thing as the volition, and names are always determined by whatever is most noble, we do not normally call it a 'passion,' but solely an 'action.'

30. The Soul Is United to All the Parts of the Body Conjointly

But in order to understand all these things more perfectly, we need to recognize that the soul is really joined to the whole body, and that we cannot properly say that it exists in any one part of the body to the exclusion of the others. For the body is a unity which is in a sense indivisible because of the arrangement of its organs, these being so related to one another that the removal of any one of them renders the whole body defective. And the soul is of such a nature that it has no relation to extension, or to the dimensions or other properties of the matter of which the body is composed: it is related solely to the whole assemblage of the body's organs. This is obvious from our inability to conceive of a half or a third of a soul, or of the extension which a soul occupies. Nor does the soul become any smaller if

we cut off some part of the body, but it becomes completely separate from the body when we break up the assemblage of the body's organs.

31. There Is a Little Gland[1] in the Brain Where the Soul Exercises Its Functions More Particularly Than in the Other Parts of the Body

We need to recognize also that although the soul is joined to the whole body, nevertheless there is a certain part of the body where it exercises its functions more particularly than in all the others. It is commonly held that this part is the brain, or perhaps the heart—the brain because the sense organs are related to it, and the heart because we feel the passions as if they were in it. But on carefully examining the matter I think I have clearly established that the part of the body in which the soul directly exercises its functions is not the heart at all, or the whole of the brain. It is rather the innermost part of the brain, which is a certain very small gland situated in the middle of the brain's substance and suspended above the passage through which the spirits in the brain's anterior cavities communicate with those in its posterior cavities. The slightest movements on the part of this gland may alter very greatly the course of these spirits, and conversely any change, however slight, taking place in the course of the spirits may do much to change the movements of the gland.

32. How We Know That This Gland Is the Principal Seat of the Soul

Apart from this gland, there cannot be any other place in the whole body where the soul directly exercises its functions. I am convinced of this by the observation that all the other parts of our brain are double, as also are all the organs of our external senses—eyes, hands, ears and so on. But in so far as we have only one simple thought about a given object at any one time, there must necessarily be some place where the two images coming through the two eyes, or the two impressions coming from a single object through the double organs of any other sense, can come together in a single image or impression before reaching the soul, so that they do not present to it two objects instead of one. We can easily understand that these images or other impressions are unified in this gland by means of the spirits which fill the cavities of the brain. But they cannot exist united in this way in any other place in the body except as a result of their being united in this gland.

33. The Seat of the Passions Is Not in the Heart

As for the opinion of those who think that the soul receives its passions in the heart, this is not worth serious consideration, since it is based solely on the fact that the passions make us feel some change in the heart. It is easy to see that the only reason why this change is felt as occurring in the heart is that there is a small nerve which descends to it from the brain—just as pain is felt as in the foot by means of the nerves in the foot, and the stars are perceived as in the sky by means of their light and the optic nerves. Thus it is no more necessary that our soul should exercise its functions directly in the heart in order to feel its passions there, than that it should be in the sky in order to see the stars there.

34. How the Soul and the Body Act on Each Other

Let us therefore take it that the soul has its principal seat in the small gland located in the middle of the brain. From there it radiates through the rest of the body by means of the animal spirits, the nerves, and even the blood, which can take on the impressions of the spirits and carry them through the arteries to all the limbs. Let us recall what we said previously about the mechanism of our body. The nerve-fibres are so distributed in all the parts of the body that when the objects of the senses produce various different movements in these parts, the fibres are occasioned to open the pores of the brain in various different ways. This, in turn, causes the animal spirits contained in these cavities to enter the muscles in various different ways. In this manner the spirits can move the limbs in all the different ways they are capable of being moved. And all the other causes that can move the spirits in different ways are sufficient to direct them into different muscles. To this we may now add that the small gland which is the principal seat of the soul is suspended within the cavities containing these spirits, so that it can be moved by them in as many different ways as there are perceptible

differences in the objects. But it can also be moved in various different ways by the soul, whose nature is such that it receives as many different impressions—that is, it has as many different perceptions as there occur different movements in this gland. And conversely, the mechanism of our body is so constructed that simply by this gland's being moved in any way by the soul or by any other cause, it drives the surrounding spirits towards the pores of the brain, which direct them through the nerves to the muscles; and in this way the gland makes the spirits move the limbs.

35. Example of the Way in which the Impressions of Objects Are United in the Gland in the Middle of the Brain

Thus, for example, if we see some animal approaching us, the light reflected from its body forms two images, one in each of our eyes; and these images form two others, by means of the optic nerves, on the internal surface of the brain facing its cavities. Then, by means of the spirits that fill these cavities, the images radiate towards the little gland which the spirits surround: the movement forming each point of one of the images tends towards the same point on the gland as the movement forming the corresponding point of the other image, which represents the same part of the animal. In this way, the two images in the brain form only one image on the gland, which acts directly upon the soul and makes it see the shape of the animal.

36. Example of the Way in which the Passions Are Aroused in the Soul

If, in addition, this shape is very strange and terrifying—that is, if it has a close relation to things which have previously been harmful to the body—this arouses the passion of anxiety in the soul, and then that of courage or perhaps fear and terror, depending upon the particular temperament of the body or the strength of the soul, and upon whether we have protected ourselves previously by defence or by flight against the harmful things to which the present impression is related. Thus in certain persons these factors dispose their brain in such a way that some of the spirits reflected from the image formed on the gland proceed from there to the nerves which serve to turn the back and move the legs in order to flee. The rest of the spirits go to nerves which expand or constrict the orifices of the heart, or else to nerves which agitate other parts of the body from which blood is sent to the heart, so that the blood is rarefied in a different manner from usual and spirits are sent to the brain which are adapted for maintaining and strengthening the passion of fear—that is, for holding open or re-opening the pores of the brain which direct the spirits into these same nerves. For merely by entering into these pores they produce in the gland a particular movement which is ordained by nature to make the soul feel this passion. And since these pores are related mainly to the little nerves which serve to contract or expand the orifices of the heart, this makes the soul feel the passion chiefly as if it were in the heart.

NOTE

1. The pineal gland.

Correspondence

Princess Elisabeth of Bohemia and René Descartes

AT 3:660[1]

Elisabeth to Descartes

[The Hague] 6 May 1643

M. Descartes,

I learned, with much joy and regret, of the plan you had to see me a few days ago; I was touched equally by your charity in willing to share yourself with an ignorant and intractable person and by the bad luck that robbed me of such a profitable conversation. M. Palotti[2] greatly augmented this latter passion in going over with me the solutions you gave him to the obscurities contained in the physics of M. Regius.[3] I would have been better instructed on these from your mouth, as I would have been on a question I proposed to that
661 professor while he was in this town, and regarding which he redirected me to you so that I might receive a satisfactory answer. The shame of showing you so disordered a style prevented me, up until now, from asking you for this favor by letter.

But today M. Palotti has given me such assurance of your goodwill toward everyone, and in particular toward me, that I chased from my mind all considerations other than that of availing myself of it. So I ask you please to tell me how the soul of a human being (it being only a thinking substance) can determine the bodily spirits, in order to bring about voluntary actions. For it seems that all determination of movement happens through the impulsion of the thing moved, by the manner in which it is pushed by that which moves it, or else by the particular qualities and shape of the surface of the latter. Physical contact is required for the first two conditions, extension for the third. You entirely exclude the one [extension] from the notion you have of the soul, and the other [physical contact] appears to me incompatible with an immaterial thing.[4] This is why I ask you for a more precise definition of the soul than the one you give in your *Metaphysics*, that is to say, of its substance separate from its action, that is, from thought.[5] For even if we were to suppose them inseparable (which is however difficult to prove in the mother's womb and in great fainting spells) as are the attributes of God, we could, in considering them apart, acquire a more perfect idea of them.

Knowing that you are the best doctor for my 662
soul, I expose to you quite freely the weaknesses of its speculations, and hope that in observing the Hippocratic oath,[6] you will supply me with remedies without making them public; such I beg of you to do, as well as to suffer the badgerings of

Your affectionate friend at your service, Elisabeth.

AT 3:663

Descartes To Elisabeth

Egmond du Hoef, 21 May 1643

Madame,

The favor with which your Highness has honored me, in allowing me to receive her orders in writing, is greater than I would ever have dared to hope; and it is more consoling to my failings than what I had hoped for with passion, which was to receive them by mouth, had I been able to be admitted the honor of paying you reverence, 664
and of offering you my very humble services when I was last in The Hague. For in that case I would have had too many marvels to admire at the same time, and seeing superhuman discourse emerging from a body so similar to those painters give to angels, I would have been delighted in the same manner as it seems to me must be those who, coming from the earth, enter newly into heaven. This would have made me less capable of responding to your Highness, who without doubt has already noticed in me this failing, when I had the honor of speaking with her before; and your clemency wanted to assuage it, in leaving me the traces of your thoughts on a paper, where, in rereading them several times and accustoming myself to consider them, I would be truly less dazzled, but I instead feel more wonder, in noticing that these thoughts not only seem ingenious at the outset, but also even more judicious and solid the more one examines them.

Excerpted from Lisa Shapiro, ed. and trans., *The Correspondence between Princess Elisabeth of Bohemia and René Descartes*, (University of Chicago Press, 2007), with the permission of University of Chicago Press.

I can say with truth that the question your Highness proposes seems to me that which, in view of my published writings, one can most rightly ask me.[7] For there are two things about the human soul on which all the knowledge we can have of its nature depends: one of which is that it thinks, and the other is that, being united to the body, it can act on and be acted upon by it.[8] I have said almost nothing about the latter, and have concentrated solely on making the first better understood, as my principal aim
665 was to prove the distinction between the soul and the body. Only the first was able to serve this aim, and the other would have been harmful to it. But, as your Highness sees so clearly that one cannot conceal anything from her, I will try here to explain the manner in which I conceive of the union of the soul with the body and how the soul has the power [*force*] to move it.

First, I consider that there are in us certain primitive notions that are like originals on the pattern of which we form all our other knowledge. There are only very few of these notions; for, after the most general—those of being, number, and duration, etc. —which apply to all that we can conceive, we have, for the body in particular, only the notion of extension, from which follow the notions of shape and movement; and for the soul alone, we have only that of thought, in which are included the perceptions of the understanding and the inclinations of the will; and finally, for the soul and the body together, we have only that of their union, on which depends that of the power the soul has to move the body and the body to act on the soul, in causing its sensations and passions.

I consider also that all human knowledge [science] consists only in distinguishing well these notions, and in attributing each of them only to those things to which it pertains. For,
666 when we want to explain some difficulty by means of a notion which does not pertain to it, we cannot fail to be mistaken; just as we are mistaken when we want to explain one of these notions by another; for being primitive, each of them can be understood only through itself. Although the use of the senses has given us notions of extension, of shapes, and of movements that are much more familiar than the others, the principal cause of our errors lies in our ordinarily wanting to use these notions to explain those things to which they do not pertain. For instance, when we want to use the imagination to conceive the nature of the soul, or better, when one wants to conceive the way in which the soul moves the body, by appealing to the way one body is moved by another body.

That is why, since, in the *Meditations* which your Highness deigned to read, I was trying to make conceivable the notions which pertain to the soul alone, distinguishing them from those which pertain to the body alone, the first thing that I ought to explain subsequently is the manner of conceiving those which pertain to the union of the soul with the body, without those which pertain to the body alone, or to the soul alone. To which it seems to me that what I wrote at the end of my response to the sixth objections can be useful;[9] for we cannot look for these simple notions elsewhere than in our soul, which has them all in itself by its nature, but which does not always distinguish one from the *667* others well enough, or even attribute them to the objects to which it ought to attribute them.

Thus, I believe that we have heretofore confused the notion of the power with which the soul acts on the body with the power with which one body acts on another; and that we have attributed the one and the other not to the soul, for we did not yet know it, but to diverse qualities of bodies, such as heaviness, heat, and others, which we have imagined to be real, that is to say, to have an existence distinct from that of body, and by consequence, to be substances, even though we have named them qualities. In order to understand them, sometimes we have used those notions that are in us for knowing body, and sometimes those which are there for knowing the soul, depending on whether what we were attributing to them was material or immaterial. For example, in supposing that heaviness is a real quality, of which we have no other knowledge but that it has the power to move a body in which it is toward the center of the earth, we have no difficulty in conceiving how it moves the body, nor how it is joined to it; and we do not think that this happens through a real contact of one surface against another, for we experience in ourselves that we have a specific notion for conceiving that; and I think *668* that we use this notion badly, in applying it to heaviness, which, as I hope to demonstrate in my Physics, is nothing really distinct from body.[10] But I do think that it was given to us for conceiving the way in which the soul moves the body.

If I were to employ more words to explain myself, I would show that I did not sufficiently recognize the incomparable mind of your Highness, and I would be too presumptuous if I dared to think that my response should

be entirely satisfactory to her; but I will try to avoid both the one and the other in adding here nothing more, except that if I am capable of writing or saying something that could be agreeable to her, I would always take it as a great honor to take up a pen or to go to The Hague for this end, and that there is nothing in the world which is so dear to me as the power to obey her commandments. But I cannot find a reason to observe the Hippocratic oath that she enjoined me to, since she communicated nothing to me that does not merit being seen and admired by all men. I can only say, on this matter, that esteeming infinitely your letter to me, I will treat it as the misers do their treasures: the more they value them the more they hide them away, and begrudging the rest of the world a view of them, they make it their sovereign good to look at them. Thus, it will be easy for me alone to enjoy the good of seeing it, and my greatest ambition is to be able to say and to be truly, Madame,

Your Highness' very humble and obedient servant, Descartes.

AT 3:683

Elisabeth To Descartes

[The Hague] 10 June 1643

M. Descartes,

Your goodwill appears not only in your showing me the faults in my reasoning and correcting them, as I expected, but also in your attempt to console me about them in order to make the knowledge of them less annoying for me. But, in detriment to your judgment, you attempt to console me about those faults with false praise. Such false praise would have been necessary to encourage me to work to remedy them had my upbringing, in a place where the ordinary way of conversing has accustomed me to understand that people are incapable of giving one true praise, not made me presume that I could not err in believing the contrary of what people speak, and had it not rendered the consideration of my imperfections so familiar that they no longer upset me more than is necessary to promote the desire to rid myself of them.

This makes me confess, without shame, that
684 I have found in myself all the causes of error which you noticed in your letter, and that as yet I have not been able to banish them entirely, for the life which I am constrained to lead does not leave enough time at my disposal to acquire a habit of meditation in accordance with your rules.[11] Now the interests of my house, which I must not neglect, now some conversations and social obligations which I cannot avoid, beat down so heavily on this weak mind with annoyance or boredom, that it is rendered useless for anything else at all for a long time afterward: this will serve, I hope, as an excuse for my stupidity in being unable to comprehend, by appeal to the idea you once had of heaviness, the idea through which we must judge how the soul (nonextended and immaterial) can move the body; nor why this power [*puissance*] to carry the body toward the center of the earth, which you earlier falsely attributed to a body as a quality, should sooner persuade us that a body can be pushed by some immaterial thing, than the demonstration of a contrary truth (which you promise in your physics) should confirm us in the opinion of its impossibility. In particular, since this idea (unable to pretend to the same perfection and objective reality as that of God) can be feigned due to the ignorance of that which truly moves these bodies toward the center, and since no material cause presents itself to the senses, one would then attribute this power to its contrary, an immaterial cause. But I nevertheless have never been able to conceive of such an immaterial thing as anything other than a negation of matter which cannot have any communication with it.

685

I admit that it would be easier for me to concede matter and extension to the soul than to concede the capacity to move a body and to be moved by it to an immaterial thing. For, if the first is achieved through *information*, it would be necessary that the spirits, which cause the movements, were intelligent, a capacity you accord to nothing corporeal.[12] And even though, in your *Metaphysical Meditations*, you show the possibility of the second, it is altogether very difficult to understand that a soul, as you have described it, after having had the faculty and the custom of reasoning well, can lose all of this by some vapors, and that, being able to subsist without the body, and having nothing in common with it, the soul is still so governed by it.

But after all, since you have undertaken to instruct me, I entertain these sentiments only as friends which I do not intend to keep, assuring myself that you will explicate the nature of an immaterial substance and the manner of its actions and passions in the body, just as well as you have all the other things that you have

wanted to teach. I beg of you also to believe that you could not perform this charity to anyone who felt more the obligation she has to you as?

Your very affectionate friend, Elisabeth.

Descartes To Elisabeth

28 June 1643, Egmond du Hoef

AT 3:690

Madame,

I have a very great obligation to your Highness in that she, after having borne my explaining myself badly in my previous letter, concerning the question which it pleased her to propose to me, deigns again to have the patience to listen to me on the same matter, and
691 to give me occasion to note the things which I omitted. Of which the principal ones seem to me to be that, after having distinguished three sorts of ideas or primitive notions which are each known in a particular way and not by a comparison of the one with the other—that is, the notion that we have of the soul, that of the body, and the union which is between the soul and the body—I ought to have explained the difference between these three sorts of notions and between the operations of the soul through which we have them, and to have stated how we render each of them familiar and easy to us. Then, after that, having said why I availed myself of the comparison with heaviness, I ought to have made clear that, even though one might want to conceive of the soul as material (which, strictly speaking, is what it is to conceive its union with the body), one would not cease to know, after that, that the soul is separable from it. That is, I think, all of what your Highness has prescribed me to do here.

First, then, I notice a great difference between these three sorts of notions. The soul is conceived only by the pure understanding [*l'entendement*]; the body, that is to say, extension, shapes, and motions, can also be known by the understanding alone, but is much better known by the understanding aided by the imagination; and finally, those things which pertain
692 to the union of the soul and the body are known only obscurely by the understanding alone, or even by the understanding aided by the imagination; but they are known very clearly by the senses. From which it follows that those who never philosophize and who use only their senses do not doubt in the least that the soul moves the body and that the body acts on the soul. But they consider the one and the other as one single thing, that is to say, they conceive of their union. For to conceive of the union between two things is to conceive of them as one single thing. Metaphysical thoughts which exercise to conceive of them as one single thing. Metaphysical thoughts which exercise the pure understanding serve to render the notion of the soul familiar. The study of mathematics, which exercises principally the imagination in its consideration of shapes and movements, accustoms us to form very distinct notions of body. And lastly, it is in using only life and ordinary conversations and in abstaining from meditating and studying those things which exercise the imagination that we learn to conceive the union of the soul and the body.

I almost fear that your Highness will think that I do not speak seriously here. But this would be contrary to the respect I owe her and that I would never neglect to pay her. And I can say with truth that the principal rule I have always observed in my studies, and that which I believe has served me the most in acquiring some bit of knowledge, is that I never spend more than a few hours each day in thoughts which occupy the imagination, and very few *693* hours a year in those which occupy the understanding alone, and that I give all the rest of my time to relaxing the senses and resting the mind; I even count, among the exercises of the imagination, all serious conversations and everything for which it is necessary to devote attention. It is this that has made me retire to the country. For even though in the most populated city in the world I could have as many hours to myself as I now employ in study, I would nevertheless not be able to use them so usefully, since my mind would be distracted by the attention the bothers of life require. I take the liberty to write of this here to your Highness in order to show that I truly admire that, amid the affairs and the cares which persons who are of a great mind and of great birth never lack, she has been able to attend to the meditations which are required in order to know well the distinction between the soul and the body.

But I judged that it was these meditations, rather than these other thoughts which require less attention, that have made her find obscurity in the notion we have of their union; as it does not seem to me that the human mind is capable of conceiving very distinctly, and at the same time, the distinction between the soul

and the body and their union, since to do so it is necessary to conceive them as one single thing and at the same time to conceive them as two, which is contradictory. On this matter (supposing your Highness still had the reasons which prove the distinction of the soul and body at the forefront of her mind and not wanting to
694 ask her to remove them from there in order to represent to herself the notion of the union that each always experiences within himself without philosophizing, in knowing that he is a single person who has together a body and a thought, which are of such a nature that this thought can move the body and sense what happens to it), I availed myself in my previous letter of a comparison between heaviness and those other qualities which we commonly imagine to be united to some bodies just as thought is united to our own, and I was not worried that this comparison hangs on qualities that are not real, even though we imagine them so, since I believed that your Highness was already entirely persuaded that the soul is a substance distinct from body.

But since your Highness notices that it is easier to attribute matter and extension to the soul than to attribute to it the capacity to move a body and to be moved by one without having matter, I beg her to feel free to attribute this matter and this extension to the soul, for to do so is to do nothing but conceive it as united with the body. After having well conceived this and having experienced it within herself, it will be easy for her to consider that the matter that she has attributed to this thought is not the thought itself, and that the extension of this matter is of another nature than the extension of this thought, in that the first is determined to a certain place, from which it excludes all other extended bodies, and this is not the case with
695 the second. In this way your Highness will not neglect to return easily to the knowledge of the distinction between the soul and the body, even though she has conceived their union.

Finally, though I believe it is very necessary to have understood well once in one's life the principles of metaphysics, since it is these that give us knowledge of God and of our soul, I also believe that it would be very harmful to occupy one's understanding often in meditating on them. For in doing so, it could not attend so well to the functions of the imagination and the senses. The best is to content oneself in retaining in one's memory and in one's belief the conclusions that one has at one time drawn from such meditation, and then to employ the rest of the time one has for study in those thoughts where the understanding acts with imagination or the senses.

The extreme devotion which I have to serve your Highness makes me hope that my frankness will not be disagreeable to her. She would have here received a longer discourse in which I would have tried to clarify all at once the difficulties of the question asked, but for a new annoyance which I have just learned about from Utrecht, that the magistrate summons me in order to verify what I wrote about one of their ministers—no matter that this is a man who has slandered me very indignantly and that what I wrote about him in my just defense was only too well known to the world—and so I am constrained to finish here, in order that I may go find the means to extricate myself as soon as I can from this chicanery.[13] I am, &c.

AT 4:1

Elisabeth To Descartes

[The Hague] 1 July 1643

M. Descartes,

I see that you have not received as much inconvenience from my esteem for your instruction and the desire to avail myself of it, as from the ingratitude of those who deprive themselves of it and would like to deprive the human species of it. I would not have sent you 2
new evidence of my ignorance until I knew you were done with those of that mindset, if Sieur Van Bergen[14] had not obliged me to it earlier, through his kindness in agreeing to stay in town, just until I gave him a response to your letter of 28 June. What you write there makes me see clearly the three sorts of notions that we have, their objects, and how we ought to make use of them.

I also find that the senses show me that the soul moves the body, but they teach me nothing (no more than do the understanding and the imagination) of the way in which it does so. For this reason, I think that there are some properties of the soul, which are unknown to us, which could perhaps overturn what your *Metaphysical Meditations* persuaded me of by such good reasoning: the nonextendedness of the soul. This doubt seems to be founded on the rule that you give there, in speaking of the true and the false, that all error comes to us in forming judgments

about that which we do not perceive well enough.[15] Though extension is not necessary to thought, neither is it at all repugnant to it, and so it could be suited to some other function of the soul which is no less essential to it. At the very least, it makes one abandon the contradiction of the Scholastics, that it [the soul] is both as a whole in the whole body and as a whole in each of its parts.[16] I do not excuse myself at all for confusing the notion of the soul with that of the body for the same reason as the vulgar; but this doesn't rid me of the first doubt, and I will lose hope of finding certitude in anything in
3 the world if you, who alone have kept me from being a skeptic, do not answer that to which my first reasoning carried me.

Even though I owe you this confession and thanks, I would think it strongly imprudent if I did not already know your kindness and generosity, equal to the rest of your merits, as much by the experience that I have already had as by reputation. You could not have attested to it in a manner more obliging than by the clarifications and counsel you have imparted to me, which I hold above all as one of the greatest treasures that could be possessed by

Your very affectionate friend at your service,
Elisabeth.

NOTES

1. I provide the reader with the volume and page references from the Adam and Tannery edition of Descartes' *Oeuvres*. The page number indicates the beginning of the page.
2. Alphonse Pollot (1602–68), whom Elisabeth refers to as Palotti, was a gentleman-in-waiting to the prince of Orange. In his letter to Pollot of October 6, 1642, Descartes notes his happiness that Elisabeth has read and seems to approve of his *Meditations*, as well as his intention to visit The Hague to meet her (see *Oeuvres de Descartes*, ed. Charles Adam and Paul Tannery, 11 vols. [Paris: Cerf, 1897–1913; new ed., Paris: Vrin, 1964–7; reprint, Paris: Vrin, 1996; cited hereafter as AT] 3, pp. 577–78). This letter would seem to mark Descartes' attempt at this meeting. Pollot's relation to Descartes began in 1638 with an exchange, through Henricus Reneri, about Descartes' *Discourse on the Method*. Pollot, as suggested here, effected the introduction between Descartes and Elisabeth. He appears to have tutored Elisabeth in geometry (see Descartes to Elisabeth, November 1645) and often served as the courier of their correspondence (see Elisabeth's letter of 24 May 1645, below). Reneri (1593–1639), a French philosopher, was a professor of philosophy at the University of Utrecht.
3. Henri le Roy or Regius (1598–1679) was a Dutch physician who took up Descartes' physics and physiology and taught them as chair of medicine at the University of Utrecht, beginning in 1638. Elisabeth's remarks here suggest that she was tutored by Regius or at least read his *Physiologia sive cognitio sanitatis* (Utrecht: Roman, 1641). While at Utrecht, beginning in 1642, Regius was attacked as promulgator of Cartesian philosophy by Professor of Theology Voetius. He was supported by Descartes in these battles until 1646. At that time there was a public falling out between Descartes and Regius upon publication of Regius' *Fundamenta Physica*. Descartes' side of this dispute can be seen in the French preface to the *Principles* and the *Comments on a Certain Broadsheet*. One can see trouble ahead in their earlier 1641 correspondence: see Descartes to Regius, May 1641 (AT 3: pp. 371–72; *The Philosophical Writings of Descartes*, ed. John Cottingham, Robert Stoothoff, and Dugald Murdoch; and in vol. 3, Anthony Kenny, 3 vols. [Cambridge: Cambridge University Press, 1984–1991, cited hereafter as CSM or CSMK, respectively], pp. 181–82), December 1641 (AT 3, pp. 454–55, CSMK, p. 199), December 1641 (AT 3, p. 460, CSMK, pp. 200–201), January 1642 (AT 3, p. 491, CSMK, pp. 491–92).
4. For a clear statement of this claim, see the Sixth Meditation argument for the real distinction of mind and body (AT 7, p. 78, CSM 2: p. 54).
5. Elisabeth here seems to be referencing the discussion in the paragraph subsequent to that containing the real distinction argument (AT 7: p. 78–80, CSM, 2: pp. 54–55), wherein Descartes details the "faculties" of extended and intellectual substances.
6. While Foucher de Careil, following Clerselier's rendering of Descartes' response, has "serment de Harpocrates" here, AT change it to Hippocrates. AT's reasoning seems sound. Not only do they follow the manuscripts, but the Hippocratic oath would have been well known to both Descartes and Elisabeth. Fabricius alludes to it, and by 1643 his work had seen more than thirty editions, one even published in Leiden in 1643 with a commentary by Meibomius. Elisabeth's later letters show her familiarity with the medical establishment, and Descartes too had interests in medicine. Moreover, while Harpocrates, or Horus, the child, is the Egyptian god of silence, and was taken up as the god of secrecy by the Greeks and Romans, there is no oath associated with him. While Harpocrates is associated with a secret medical profession in certain monuments, this same secret is contained in the Hippocratic oath: "About whatever I may see or hear in treatment, or even without treatment, in the life of human beings—things that should not ever be blurted out outside—I will remain silent, holding such things to be unutterable [sacred, not to be divulged]." Translation by Heinrich Von Staden, "In a Pure and Holy Way: Personal and Professional Conduct in the Hippocratic Oath," *Journal of the History of Medicine and Allied Sciences* 51 (1996): pp. 406–8.
7. At this point, Descartes had published the *Discourse on the Method*, with accompanying essays (1637),

and the *Meditations*, along with Objections and Replies (1641, 1642). He says little in those works about the philosophical basis of mind-body interaction. Gassendi, in the Fifth Objections, had raised a similar question, though he met with a much less hospitable reply. See AT 7, pp. 343–44, 389–90, 9:2 p. 13; CSM 2, pp. 238–39, 266, 275–76.

8. *Agir et patir avec lui:* In English, it is difficult to bring out the parallel between active and passive, which preserves the tie to the passions of the soul that will figure prominently in the later correspondence.
9. AT 7, pp. 444–45; CSM 2, pp. 299–300.
10. *Principles* 4.20–27 (AT 7A, pp. 212–16; CSM 1, pp. 268–70).
11. Elisabeth here seems to be referring to what Descartes writes in the preface to reader of the *Meditations*, and in the postulates of the geometrical exposition of his philosophy in the Second Replies, where he requires that his readers "meditate seriously with me, and withdraw their minds from the senses and from all preconceived opinions" (AT 7, p. 9; CSM 2, p. 8; see also AT 7, p. 162ff.; CSM 2, p. 114ff.). Doing so, however, requires that one be able to "expressly rid [one's] . . . mind of all worries and arrange for [oneself] . . . a clear stretch of free time," as the meditator does in the First Meditation (AT 7, p. 18; CSM 2, p. 17). It is this luxury Elisabeth cannot afford.
12. I have here retained the French *information.* It is hard to determine what theoretical model Elisabeth is adverting to. On the one hand, it is tempting to think that she is invoking the Aristotelian doctrine that the soul is the form of the body and so informs the body. On the other hand, her concern with the intelligence of corporeal spirits suggests that she is referring to a Stoic account of cognitive faculties and intentional action. The Stoics explained the cohesion of bodies and their motions toward some end, as well as the rational faculties Descartes accords to the soul (and so, one might say, the information of substances), by appeal to that part of matter termed *pneuma.*
13. See the Letter to Voetius, AT 8B, pp. 3–194. Parts of this very long letter are translated in CSMK, pp. 220–24. This letter, which was published in Latin and simultaneously in Flemish translation in May 1643, was written as a reply to the pointed published attacks on Cartesianism by Voetius. Voetius, as rector of the University of Utrecht, had earlier arranged for the formal condemnation of Cartesian philosophy at the university. For further reading on this dispute, see Verbeek and Marion, *La querelle d'Utrecht*; and Verbeek, *Descartes and the Dutch.* See also Descartes to Father Dinet, especially AT 7, p. 582ff.; CSM 2, p. 393ff., and the postscript of Elisabeth's letter of June 22, 1645, below.
14. Anthonie Studler Van Surck, sieur de Bergen (1606–66), was Descartes' banker in Holland and sometimes acted as intermediary for Descartes' letters. In particular he often served as intermediary in Descartes' correspondence with Huygens. Elisabeth might well have known him through this connection with Huygens, since she too corresponded with Constantijn Huygens (1596–1687), a noted humanist scholar and father of the mathematician and physicist Christian Huygens (1629–95). In addition, the sieur de Bergen was charged with the distribution of the *Principles* in Holland, while Descartes was in France in 1644.
15. See the rule arrived at and articulated in the Fourth Meditation: "If, however, I simply refrain from making a judgment in cases where I do not perceive the truth with sufficient clarity and distinctness, then it is clear that I am behaving correctly and avoiding error" (AT 7, p. 59; CSM 2, p. 41).
16. See, for example, Aquinas, *Summa Theologica*, I, q.76 a.8.

The Akan Concept of a Person

Kwame Gyekye

Introduction

A number of scholars, including philosophers, tend to squirm a little at the mention of 'African philosophy,' though they do not do so at the mention of African art, music, history, anthropology, religion, etc. While the latter cluster of disciplines is being cultivated or pursued in the various centres or institutes of African studies in universities round the world, African philosophy as such is relegated to limbo because it is considered to be nonexistent. Philosophy is thus assumed to be a special relish of the peoples of the West and the East. To a very great extent the lack of writing in Africa's historical past, leading in turn to the absence of a doxographic tradition, that is, a tradition of recorded opinions, has been responsible for the assumption that there is no such thing as African philosophy.

We do not ask the question whether there is European philosophy or Greek philosophy simply because there are the classic *Dialogues*, *Treatises*, *Essays*, *Philosophical Investigations*, which one can immediately delve into if he wants to study European or Greek philosophy. In Africa, traditionally, there has been a dearth of such philosophical classics. Yet this fact does not in any way argue the nonexistence of African philosophy. For it is known that Socrates, the celebrated ancient Greek philosopher, did not write anything, although he inherited a written culture; but it is also known that he *philosophized*. In India 'the Upanishads which are imbued with philosophy . . . were not written down for centuries.'[1] An eminent Indian philosopher wrote: 'The Vedas were handed down from mouth to mouth from a period of unknown antiquity. . . . When the Vedas were composed, there was probably no system of writing prevalent in India.'[2] (The Vedas constitute the religious and philosophical classics of India. The Upanishads form the concluding portions of the Vedas.) And I learn that Buddha, the ancient Indian philosopher and religious thinker, 'wrote no book, but taught orally.'[3] Thus African philosophy is none the worse for the absence, traditionally, of written philosophical literature. To deny to African peoples philosophical thought is to imply that they are unable to make philosophical sense of, or to conceptualize, their experiences; it is in fact to deny them their humanity. For philosophy of some kind is behind the thought and action of every people. It constitutes the intellectual sheet anchor of their life in its totality.

African philosophic thought not only forms part of the oral literature of the peoples; it is also expressed or reflected in real and vital attitudes. In Africa a great deal of philosophical material is embedded in the proverbs,[4] myths and folk tales, folk songs, rituals, beliefs, customs, and traditions of the peoples. The interested and careful philosopher can perceive the philosophical relevance of such material and may come across ideas or doctrines or problems that may have some affinity with those of the West or the East, but which originated from the peoples themselves.

After these dialectical preambles, I wish now to turn to a discussion of the Akan[5] concept of a person, in which I shall attempt to interpret, reconstruct, and sort out in a more sophisticated way the elements of the Akan collective thought on the nature of a person, and provide the necessary conceptual or theoretical trimming such as is required by the anthropological and sociological accounts.

I. *OKRA* (SOUL)

We are given to understand from anthropological accounts that the Akans hold a tripartite conception of a person, considering a human being to be constituted by three elements: *okra*, *ṣunsum*, and *honam* (or *nipadua*: body).

The *okra* is considered to be that which constitutes the very inner self of the individual, the principle of life of that individual, and the embodiment and transmitter of his destiny (fate:

Gyekye, Kwame (1978). "The Akan Concept of a Person." *International Philosophical Quarterly* 18(3): 277–287.

nkrabea). It is thought to be a spark of God (*Onyame*) in man. It is thus divine and has an ante-mundane existence with God; it derives directly from God. The *okra*, therefore, might be considered as the equivalent of the concept of the soul in other metaphysical systems.

The presence of this divine principle in a human being may have been the basis of the Akan proverb *Nnipa nyinaa ye Onyame mma, obiara nye asase ba* ('All men are the children of God; no one is a child of the Earth').

The conception of the *okra* as the life principle in a person, his vital force, the source of his energy, is linked closely with another concept, namely *honhom*. *Honhom* means 'breath'; it is the noun form of *home*, to breathe. When a man is dead it is said *ne honhom ko* ('his breath is gone') or *ne 'kra afi ne ho* ('his soul has withdrawn from his body'). The two sentences, one with *honhom* as subject and the other with *okra* as the subject, do, in fact, say the same thing; they express the same thought, the death of the person. The departure of the soul from the body means the death of a person, and so does ceasing to breathe. Yet this does not mean that the *honhom* (breath) is the *okra* (soul). The *okra* is that which 'causes' the breathing. Thus, the *honhom* is the tangible manifestation or evidence of the *okra*. (I must say, however, that in some dialects of the Akan language *honhom* has come to be used interchangeably with *sunsum*, so that the phrase *honhom bone* has come to mean the same thing as *sunsum bone*, i.e. evil 'spirit.' The identification of the *honhom* with the *sunsum* seems to me to be a recent idea and may have resulted from the translation of the Bible into the various Akan dialects: *honhom* must have been used to translate the Greek *pneuma*, breath, spirit. The clarification of the concepts of *okra*, *honhom*, and *sunsum* (spirit) is the burden of this paper.)

II. *SUNSUM* (SPIRIT)

Sunsum is another of the constituent elements of a person. It has usually been rendered in English as 'spirit.' In some of the literature on Western metaphysics 'spirit' appears to be a generic or comprehensive concept under which are subsumed specific concepts such as soul, mind, self, consciousness—all of which are, however, considered to be identical. But some Western philosophers distinguish the mind from the soul, for while they are prepared to admit that a human being has a mind (which they would identify with the brain or a brain state), they deny the existence of the soul mainly because of the immortality attribute that has traditionally been claimed for it.[6] In the Akan metaphysics of the person, however, 'spirit' is a specific concept. (I shall show in a later publication on Akan ontology that the concept is also used generically in other contexts.) It appears from the anthropological accounts that even when it is used as a specific concept, 'spirit' (*sunsum*) is not identical with the soul (*okra*) as they do not refer to the same thing. However, the anthropological accounts of the *sunsum* involve some conceptual blunders, as I hope to show presently. As for the mind (when it is not identified with the soul) it might be rendered also by *sunsum*, judging from the functions that are attributed by the Akans to the latter (see below).

On the surface it might appear that 'spirit' is not an appropriate rendition for *sunsum;* but after clearing some misconceptions engendered by some anthropological writings, I shall show that it is an appropriate rendition but that its real nature requires some clarification. Anthropologists and sociologists have held (i) that the *sunsum* derives from the father,[7] (ii) that it is not divine,[8] and (iii) that it perishes with the disintegration of the *honhom*,[9] that is, the material component of a person. It seems to me, however, that all these three characterizations of the *sunsum* are incorrect.

Let us first take up the third characterization of the *sunsum*, namely, that it is something that perishes with the perishing of the body. Now, if a body, a physical object, perishes along with the *sunsum*, then it would follow that the *sunsum* also is something physical or material. As a matter of fact, Danquah in his philosophical analysis concludes that '*sunsum* is, in fact, the matter or the physical basis of the ultimate ideal of which *okra* (soul) is the form and the spiritual or mental basis.'[10] Elsewhere he speaks of an 'interaction of the material mechanism (*sunsum*) with the soul,' and assimilates the *sunsum* to the 'sensible form' of Aristotle's metaphysics of substance and the *okra* to the 'intelligible form.'[11] One would conclude from these statements that Danquah also conceived the *sunsum* as material (although some other statements of his would seem to contradict this). (*See below.*) The relationship between the *honam* (body) and the *sunsum* (supposedly bodily), however, is left unexplained. Thus philosophical, sociological, and anthropological accounts of the nature of a person have given us the

impression that the Akans held a triparite conception of a human being:

okra (soul)	—	immaterial
sunsum ('spirit')	—	material (?)
honam (body)	—	material

As we shall see presently, however, this account or analysis of a person, particularly the characterization of the *sunsum* ('spirit') as material, is not satisfactory. I must admit at this point that the real nature of the *sunsum* presents some difficulty for the Akan metaphysics of a person and has been a source of confusion for scholars. The difficulty is not insoluble, however.

There are many things said regarding the functions or activities of the *sunsum* which indicate that it surely is neither material (physical), nor mortal, nor derived from the father. Busia says that the *sunsum* 'is what moulds the child's personality and disposition. It is that which determines his character and individuality.'[12] Danquah says: 'But we now know the notion which corresponds to the Akan 'sunsum,' namely, not 'spirit' as such but 'personality' which covers the relation of the 'body' to the 'soul' (*Okra*).'[13] That the *sunsum* constitutes the personality and character of a person is stated by Danquah in several pages of his book.[14] Rattray also observed that the *sunsum* is the basis of character and personality.[15] There are indeed some sentences in the Akan language in which the expression *sunsum* is used in obvious reference to personality (or qualities or traits in a person's character). Thus, for 'he has a strong personality' the Akans would say *'ne sunsum ye duru'* (i.e. his *sunsum* is 'heavy' or 'weighty'). When a man is generous they say that he has a good *sunsum* (*owo sunsum pa*). When a man has an impressive or imposing personality they say that he has an overshadowing *sunsum* (*ne sunsum hye me so*). In fact, sometimes in describing a dignified person they would simply say, '*owo sunsum*,' that is, he has a commanding presence. And a man may be said to have a 'gentle' *sunsum*, a 'forceful' *sunsum*, a 'submissive' or 'weak' *sunsum*. Thus, the concept of the *sunsum* would correspond in some ways to what is meant by personality, as was observed by some earlier investigators.

Thus, it is now clear that in Akan conceptions the *sunsum* ('spirit') is the basis of a man's personality, his distinctive character and, in the words of Busia, 'his ego.'[16] Personality, of course, is a word that has been given various definitions by psychologists. But I believe that whatever else that concept may involve, it certainly involves the idea of a set of characteristics as shown in a person's behavior—his thoughts, feelings, actions, etc. (I do not think that it refers exclusively to a person's physical appearance.) Thus, if the *sunsum* is that which embodies a man's personality, it just cannot be a physical thing, for qualities of personality such as courage, generosity, jealousy, gentleness, forcefulness, meekness, and dignity are not sensible or physical qualities; they are *psychical*. The jealous man *feels* ill or unhappy because of a possible or actual loss of position, status, expectations, or because of the better fortune of others; a courageous man is able to control *fear* in the face of danger, pain, misfortune, etc.; the ambitious man has strong *desire* to achieve something. The expressions *feel*, *fear*, and *desire* are of course psychical (psychological), not physicalistic, expressions. (In Akan metaphysics there is no room for materialism, the doctrine held by some philosophers in the West that a person is fundamentally a physical entity and that what is referred to as mind or soul is in fact identifiable with a person's brain, which is a physical organ.) Thus, if in fact personality is the function of the *sunsum*, then the latter cannot conceptually be held to be physical or material; it must surely be something (*ade*) immaterial, i.e. spiritual.

We have already noted certain statements of Danquah which suggest a physicalistic interpretation of the *sunsum*. On the other hand, he also maintains that 'it is the *sunsum* that experiences,'[17] and that it is through it that 'the *okra* or soul manifests itself in the world of experience.'[18] Elsewhere he says of the *sunsum*: 'It is the bearer of conscious experience, the unconscious or subliminal self remaining over as the okra or soul.'[19] It is not clear what Danquah means by the 'bearer' of experience. Perhaps what he means is that the *sunsum* is the subject of experience; that which experiences.

This being so, I would think, at least provisionally, that the subject of experience cannot be physical. If, as he thought,[20] it is the *sunsum* which makes it possible for the destiny (*nkrabea*: fate) of the soul to be 'realized' or 'carried out' on earth, then, like the *okra* (soul), an aspect of whose function it was going to perform, the *sunsum* also must be considered as something spiritual, not physical. Danquah's position on the concept of the *sunsum* is ambivalent. And

so is Busia's. Busia says that one part of a man is 'the personality that comes indirectly from the supreme Being,'[21] that is, God. By 'personality' Busia must, on his own showing,[22] be referring to the *sunsum* of a man, which must, according to my analysis of that concept, derive directly from God, and not from the father. It must, therefore, be divine and immortal, contrary to what he and others thought.

The explanation the Akans give of the phenomenon of dreaming also indicates that the *sunsum* is something spiritual. For the Akans, as for Sigmund Freud, dreams are not somatic but psychic phenomena. They believe that in a dream it is the person's *sunsum* that is the 'actor.' In sleep the *sunsum* is said to be released from the fetters of the body. It, as it were, fashions for itself a new world of forms with the materials of its waking experience. Thus, although the person is deeply asleep, his body (*honam*) lying in bed, yet he may 'see' himself standing on the top of a mountain or driving a car or fighting with someone. The actor in any of these actions is thought to be the *sunsum*, which thus can leave the body and return to it.

As the basis or determinant of personality traits—which are nonsensible—as a co-performer of the activities or functions of the *okra* (soul), undoubtedly thought to be a spiritual entity, and as the *dramatis persona* of the spiritual or psychical phenomenon of dreaming, the *sunsum* must be something spiritual (immaterial). This is the reason for my earlier assertion that 'spirit' might not be an inappropriate translation for *sunsum*, that is to say, the *sensum* is something spiritual. On my analysis, then, we would have the following picture:

Okra (soul)	—	immaterial
Sunsum (spirit)		(spiritual)
Honam (body)	—	material (physical)

Thus, the Akans hold a dualistic conception of a person: a person is constituted by two principal substances, one spiritual (immaterial) and the other physical (material).

III. Relation of *OKRA* and *SUNSUM*

Now having shown that the *sunsum* is in fact something spiritual (and for this reason I shall henceforth use the word 'spirit' or 'spiritual' in reference to *sunsum* without quotes), we must go on to examine whether the expressions *sunsum* (spirit) and *okra* (soul) denote the same object in Akan metaphysics and philosophical psychology. In the course of my field research I was informed by a number of elderly people I interviewed that the *sunsum*, *okra*, and *honhom* ('breath') are identical; it is one entity that goes under three names. I have already shown that while there is a close link between the *okra* and the *honhom*, the two cannot, nevertheless, be identified.[23] What about the *sunsum* and the okra? Can they be identified?

To say that the two can be identified would logically mean that whatever can be asserted of one can or must be asserted of the other. Yet there are some things the Akans say about the *sunsum* which are not said of the *okra;* the predicates or attributes of the two are different. Thus, the need for a reconstruction of the relation between the *okra* and the *sunsum*. The Akans say:

1) *ne 'kra di awerehow* ('his *kra* is sad'; never, 'his *sunsum* is sad').
2) *ne 'kra teetee* ('his *kra* is worried or disturbed').
3) *ne 'kra adwane* ('his *kra* has run away,' an expression they use when someone is scared to death).
4) *ne kra ye* ('his *kra* is good,' a sentence they use when they want to say that a person is lucky or fortunate).
5) *ne kra afi ne ho* ('his *kra* has withdrawn from his body').
6) *ne kara dii n'akyi, anka owui* ('but for his *kra* that followed him, he would have died').
7) *ne kra aniagye* ('his *kra* is happy').

In all such statements, the attributions are made to the *okra* (*kra*: soul), never to the *sunsum.* On the other hand, the Akans say:

1) *owo sunsum* ('he has *sunsum*,' an expression they use when they want to refer to someone as dignified and as having a commanding presence. Here they never say *owo* (*o*) *kra* ('he has *okra*,' soul, for it is believed that every human being has a soul, the principle of life, but the nature of the *sunsum* differs from person to person; thus they speak of 'gentle *sunsum*,' 'forceful *sunsum*,' 'weak or strong *sunsum*,' etc.).
2) *ne sunsum ye duru* ('his *sunsum* is heavy or weighty,' i.e. he has a strong personality).

3) *ne sunsum hye* (or *to*) *me so* ('his *sunsum* overshadows mine').
4) *obi sunsum so kyen obi dee* ('someone's *sunsum* is bigger or greater than another's').
5) *owo sunsum pa* ('he has a good spirit,' i.e. he is a generous person).

In all such statements the attributions are made to the *sunsum*, never to the *okra*.

Now, given *x* and *y*, if whatever is asserted of *x* can be asserted of *y*, then *x* can be said to be identical with *y*. If there is at least one predicate, which *x* has but *y* does not have, then *x* and *y* are not identical. On this showing, to the extent that things that are asserted or predicated of the *okra* do not apply to the *sunsum*, the two cannot logically be identified. But while they are logically and functionally distinct, they are not *ontologically* distinct. That is to say, they are not separate existences held together by an external bond. They are a unity in duality, a duality in unity. The distinction is not a relation between two independent entities. And the *sunsum* may, perhaps more accurately, be characterized as *a state* of the *okra* (soul). As mentioned earlier, the *okra* is the principle of life of a person and the embodiment and transmitter of his destiny (*nkrabea*). Personality and character traits of a person are the function of the *sunsum*. The *sunsum* appears to be the source of dynamism of a man, the really active part or force of the psychological system of man. It is said to have extra-sensory powers; it is that which thinks, desires, etc. It is not in any way identical with the brain. Rather it acts upon the brain (*adwen*); it is that which makes the *adwen* (brain) work. In short, it is upon the *sunsum* that man's health, worldly power, influence, position, success, etc. would depend.

Moreover, moral predicates are generally ascribed to the *sunsum*. Lystad is, thus, wrong when he says: 'In many respects the *sunsum* or spirit is so identical with the *okra* or soul in its functions that it is difficult to distinguish between them.'[24]

In the Akan conception of a person, the soul (*okra*) is held to be a mental or spiritual entity (substance). It is not a bundle of qualities or perceptions, as is held in some Western philosophies. The basis for this assertion is the Akan belief in disembodied survival. A bundle theory of substance implies the elimination of the notion of substance, for if a substance is held to be a bundle or collection of qualities or perceptions it would mean that when the qualities or perceptions are removed nothing would be left; there would then be no substance, i.e. no substratum or 'owner' of those qualities. Thus, if the soul or mind is held to be a bundle of perceptions, as in Hume, it would be impossible to talk of disembodied survival in the form of a soul or self since the bundle itself is an abstraction. One Akan maxim, expressed epigrammatically, is that 'when a man dies he is not (really) dead' (*onipa wu a na onwui*). What they imply by this is that there is something in a human being which is eternal and indestructible, and which continues to exist in the world of spirits (*asamandow*). An Akan motif expresses the following thought: 'Could God die, I will die' (*Onyame bewu na m'awu*). In Akan metaphysics God is held to be eternal, immortal (*Odomankoma*), and what is being asserted in the above thought is that since God will not die, a person, that is, his *okra* (soul), conceived as a spark of God in a person, will not die either. That is to say, the soul of man is immortal. But—and this is the point I want to make—the attributes of immortality and eternity make sense if, and only if, the soul is held to be a substance, and not a bundle of qualities or perceptions.

But where in a human being is this mental or spiritual substance located? Descartes thought that the soul is in the pineal gland. The Akans also seem to hold that the soul (*okra*) is lodged in the head of a person, although they do not mention any specific part of the head where it is. But although it is in the head, 'you cannot see it with your natural eyes,' as they would put it, since it is an immaterial substance.

That the soul is in the head (*eti*, *ti*), may be inferred from the following expressions of the Akans: When they want to say that a person is lucky or fortunate they would say *ne ti ye* ('his *head* is well') or *ne 'kra ye* ('his soul is well'). Both sentences express the same thought. And when a person is constantly afflicted with misfortunes he would say '*me ti nnye*' ('my *head* is not well') or '*me 'kra nnye*' ('my *soul* is not well'). It may be inferred from such expressions that there is some kind of connection between the head and the soul. And although they cannot point to a specific part of the head as the 'residence' of the soul, it may be conjectured that it is in the region of the brain (*adwen*), which, as stated earlier, receives its activism from the *sunsum* (spirit), a state of the soul (*okra*). That is, the mind (or, soul) acts on the brain in a specific locality, not that it is itself actually localized.

The Akan conception of a person, as it appears in my analysis, is thoroughly dualistic, not tripartite. A dualistic conception of a person does not necessarily carry with it a belief in a causal relation or interaction between the two parts of the person, soul and body. For instance, some dualistic philosophers in the West maintain a doctrine of psycho-physical parallelism, which completely denies causal interaction between body and soul. Others, also dualists, maintain a doctrine of epiphenomenalism which, while not completely rejecting causal interaction, holds that the causal direction goes in one way only, namely, from body to mind; such a doctrine is thus not an interactionist doctrine. The Akans, however, maintain a thorough interactionist position on the relation between soul and body. They believe that not only does the body have a causal influence on the soul but also the soul has a causal influence on the body (*honam*). What happens to the soul (*okra*) takes effect or reflects on the condition of the body. Similarly, what happens to the body reflects on the condition of the soul.

It is the actual bodily or physical behavior of a person that gives some idea of the condition of the soul. Thus, if the physical behavior of a person suggests that he is happy they would say *ne 'kra ani agye* ('his soul is happy'); if unhappy or morose they would say, *ne 'kra di awerchow* ('his soul is sorrowful'). When the *okra* (soul) is enfeebled or injured by evil spirits ill health results; and the poor conditions of the body also affect the condition of the soul. That is, the condition of the soul depends on the condition of the body. As a matter of fact, the belief in psycho-physical causal interaction is the whole basis of spiritual or psychical healing. There are certain diseases which are believed to be 'spiritual diseases' (*sunsum yare*) and cannot be healed by the application of physical therapy. In such diseases attention is paid to both physiological and spiritual aspects of the person. Unless the soul is healed, the body will not respond to any physical treatment. The removal of a disease of the soul is the activity of the diviners or the 'medicine men' (*adunsifo*).

Some similarities have been discovered between the functions and activities of the *sunsum* of the Akan psychology and the ego of Freud. An essential task of the ego is to engage in intercourse with the external world. Like the *sunsum*, it directs the business of everyday living; it is the executive of the personality, that is, the psychological system. It is the representative of the id in the external world. An aspect of the nature of the *sunsum* is or may be similar to the ego. The *sunsum* of the Akan psyche is not always conscious, and a man does not always know what his *sunsum* wants. It is believed that it is the *sunsum* that the Akan diviner (*okomfo*), believed to possess extrasensory abilities, communicates with. It tells the diviner what it really wants without the person knowing or being aware of what he wants; thus, the *sunsum* may be unconscious. Freud said: 'And it is indeed the case that large portions of the ego and super-ego can remain unconscious and are normally unconscious. That is to say, the individual knows nothing of their contents and it requires an expenditure of effort to make them conscious.'[25] It is, I suppose, for these reasons that some scholars[26] have not hesitated to identify the *sunsum* with the ego of Freud, and having done so go on to identify the *okra* with the id.

But there are dissimilarities which must be stated. Firstly, in Freud the id is the original system of the psyche, the matrix within which the ego and the super-ego become differentiated. But in the Akan conception both the *okra* and the *sunsum* at once constitute the original system of the psyche. Unlike the id, the *okra* is not the only component that is present at birth. Secondly, in Freud the ego and the super-ego are formed or developed later. In Akan the *sunsum* is not formed later; it was part and parcel of the original psychical structure, the *okra*, soul. At birth the child possesses a *sunsum*, just as it at that time possesses an *okra*. Freud thought in fact that the mental structure of a man was pretty well formed by the end of the fifth year. Thirdly, the super-ego is the moral dimension of personality; it represents the claims of morality.[27] In the Akan system, as stated earlier, moral attributes are generally attributed to the *sunsum*. Thus the *sunsum* of the Akan seems to perform aspects of the functions of both the ego and the super-ego of Freud.

It seems to me that an interactionist psycho-physical dualism is more realistic than materialism, epiphenomenalism, parallelism, etc. Even apart from the prospects for disembodied survival which this theory of a person holds out, it has had significant pragmatic consequences in Akan communities as evidenced in the applications of actual psycho-physical therapies. There are countless testimonies of people who have been subjected to physical treatment for months or years in modern hospitals without being cured, but who have actually been healed by traditional 'medicine men' applying both physical and psychical (spiritual) methods.

All this seems to underline the facts that a human being is not just a bag of flesh and bones, that he is a complex being who cannot completely be explained by the same laws of physics used to explain inanimate things, and that our world with all its complex and strange phenomena cannot simply be reduced to physics.

NOTES

1. G. Parrinder. *Religion in Africa* (London: UK, 1969), p. 25.
2. S. Dasgupta. *A History of Indian Philosophy* (Cambridge, UK, 1963), 1. 10.
3. P. T. Raju. *The Philosophical Traditions of India* (London: Allen & Unwin, 1971), p. 114.
4. See my "The Philosophical Relevance of Akan Proverbs," in *Second Order, An African Journal of Philosophy*, 4:2 (July 1975).
5. The Akans constitute about two-thirds of the peoples of Ghana.
6. See, e.g., Jenny Teichman. *The Mind and the Soul* (London: Routledge, 1974), 3f.
7. K. A. Busia, "The Ashanti of the Gold Coast," in D. Forde, *African Worlds* (London: Oxford Univ. Press, 954), pp. 197 and 200; M. Fortes, *Kingship and the Social Order* (Chicago: Aldine, 1969), p. 199, note 14; R. A. Lystad, *The Ashanti: A Proud People* (New Brunswick, New Jersey; Rutgers Univ. Press, 1958), p. 155; Rev. Peter Kwasi Sarpong, *Ghana in Retrospect: Some Aspects of the Ghanian Culture* (Accra-Tema: Ghana Publishing Corp., 1974), p. 37.
8. Busia, "The Ashanti of the Gold Coast"; Lystad, *The Ashanti: A Proud People;* S. L. R. Meyerowith, *The Sacred State of the Akan* (London: Faber & Faber, 1949) 86; "Concepts of the Soul in Akan," in *Africa* (Cambridge, UK: Cambridge Univ. Press, 1951), p. 26.
9. Busia, "The Ashanti of the Gold Coast"; Lystad, *The Ashanti: A Proud People*; P. A. Twumasi, *Medical Systems in Ghana: A Study in Medical Sociology* (Ghana Publishing Corp., 1975), p. 22.
10. J. B. Danquah. *The Akan Doctrine of God* (London: J. B. Danquah, 1944), p. 115.
11. Ibid., p. 116.
12. Busia, "The Ashanti of the Gold Coast," p. 197.
13. Danquah. *The Akan Doctrine of God*, p. 66.
14. E.g., pp. 67, 75, 83, 205.
15. R. S. Rattray. *Ashanti* (Oxford UK: Clarendon Press, 1923), p. 46.
16. Busia, "The Ashanti of the Gold Coast," p. 197.
17. Danquah, op. cit., 67.
18. Danquah, loc. cit.
19. Danquah, *The Akan Doctrine of God*, p. 112.
20. Ibid., pp. 66–67, 115.
21. Busia, "The Ashanti of the Gold Coast," p. 200.
22. Busia, loc. cit. in note 16.
23. This work, sec p. 278 above.
24. Lystad, *The Ashanti: A Proud People*, p. 158.
25. Sigmund Freud. *New Introductory Lectures on Psycho-analysis* (London: Penguin, 1973), pp. 101–102.
26. See, for instance, Meyerowitz, *The Sacred State of the Akan*, p. 84; Kwame Gyekye, "Concepts of the Soul in Akan," p. 26; Rev. H. Debrunner, *Witchcraft in Ghana* (Kumasi: Presbyterian Book Depot, 1959), p. 15.
27. Freud. *New Introductory Lectures on Psycho-analysis*, p. 92.

The Floating Man

Avicenna

. . . For the purposes of establishing the existence of the soul belonging to us, here we have to provide a pointer that serves [both] as alert and reminder by hitting the mark with anyone who is at all capable of catching sight of the truth on his own, and also does not require straightening out his way of thinking, or hitting him over the head with it, or steering him away from sophisms. So we say that it has to be imagined as though one of us were created whole in an instant but his sight is veiled from directly observing the things of the external world. He is created as though floating in air or in a void but without the air supporting him in such a way that he would have to feel it, and the limbs of his body are stretched out and away from one another, so they do not come into contact or touch. Then he considers whether he can assert the existence of his self. He has no doubts about asserting his self as something that exists without also [having to] assert the existence of any of his exterior or interior parts, his heart, his brain, or anything external. He will, in fact, be asserting the existence of his self without

Excerpts from Avicenna (Ibn Sina), *The Cure*, approx. 1027. Translated in Jon McGinnis and David C. Reisman, eds. and trans., *Classical Arabic Philosophy: An Anthology of Sources* (Hackett, 2007). Reprinted with the permission of the translator and the publisher.

asserting that it has length, breadth, or depth, and, if it were even possible for him in such a state to imagine a hand or some other extremity, he would not imagine it as a part of his self or as a necessary condition of his self—and you know that what can be asserted as existing is not the same as what cannot be so asserted and that what is stipulated is not the same as what is not stipulated. Thus, the self whose existence he asserted is his unique characteristic, in the sense that it is he himself, not his body and its parts, which he did not so assert. Thus, what [the reader] has been alerted to is a way to be made alert to the existence of the soul as something that is not the body—nor in fact *any* body—to recognize it and be aware of it, if it is in fact the case that he has been disregarding it and needed to be hit over the head with it. . . .

4. . . . Let us return to what was stated earlier on our part. We say: If a human were created in a single instant such that his limbs were separated from one another and he could not see them, and it happened that he could not feel them and they did not touch one another and he could not hear a single sound, he would not know that any of his organs exist, but he would know that he exists as uniquely a single thing despite not knowing everything else. However, what is unknown is not the same as what is known! These bodily members that we have are really only just like clothes that, because they have always been associated with us, we have come to think of as parts of ourselves. When we imagine ourselves, we do not imagine them bare; rather, we imagine [ourselves] to have enveloping bodies. The reason for that is the permanent association [of the two]. The fact, however, is that we have become accustomed to stripping off and discarding clothes in a way we are not accustomed to doing with the bodily members, and so our belief that these are parts of us is more firmly entrenched than our belief that our garments are parts of us.

5. If it is . . . that such a body is not the whole body but rather one specific bodily organ, then that organ would be the thing that I believe to be me—unless what is intended in my believing that it is me is not that organ, even if it must have that organ.[1] If, however, what that organ is, namely, its being a heart, a brain, or some other organ or organs with this description, is identical to it or its totality is identical to the thing that I perceive to be myself, then my perception that I am must be my perception of that thing. But one thing from a single perspective cannot be both what is perceived and other than what is perceived.[2] The situation is not like that anyway; for it is rather by sensing, listening, and experiential knowledge that I know that I have a heart and a brain, not because I know that I am I. Thus, that organ on its own would not be the thing that I perceive to be me essentially but only me accidentally, whereas the aim in knowing about myself that I am me (that is, the aim that I intend when I say '*I* sensed, *I* intellected, *I* acted, and *I*, as something different than these descriptions, joined them together') is what I call 'I.'

6. Now, if someone said, 'You also do not know that [the 'I'] is a soul,' I would say that I *always* know it as the thing intended by what I call the 'soul.' I might not know it by the term 'soul,' but once I understand what I mean by soul, I understand that it is that thing and that it is what uses [bodily] instruments such as the motive and perceptive faculties. It is only as long as I do not understand the meaning of 'soul' that I do not recognize [that]. That is not the case with the heart or the brain; for I may understand what is meant by 'heart' and 'brain,' but I do not know that [they are the 'I']. When I mean by 'soul' that it is the thing that is the principle of these motions and perceptions that I have and is what these [motions and perceptions] are traced back to in this whole, I recognize that either it is in actual fact the 'I' or it is the 'I' as something using this body. Then, it would be as though I now am unable to distinguish the perception of me as distinct from the mixed perception that there is something that uses the body, and that there is something that is joined with the body.

7. As for whether it is a body or not a body, in my opinion it is by no means necessary that it be a body, nor that it appear to me in imagined form as any body whatsoever. Instead, its imagined form appears to me to be precisely *without* any corporeality. So I will have understood some part of the aspect of its not being a body when I do *not* understand it to have any corporeality at the very same time that I understand [what it is]. Then, when I undertake an independent verification, the more I add corporeality to this thing that is the principle of these acts, the less conceivable it will be for that thing to be a body. How much more fitting it would be for its first representation in my soul to be something that is different from these exterior aspects, and I am then misled by the association with bodily instruments, the sensory observation of those, and the issuance of actions from them, and I believe that [those exterior aspects] are like parts of me. It is not when an error has been made about something that a judgment must pertain

to it, but rather when the judgment pertains to what it is that has to be intellected. And it is not when I am investigating whether it exists and whether it is not a body that I am wholly ignorant of [these questions], but rather when I neglect [to consider these questions]. It is often the case that knowledge about something is close at hand but one overlooks it, and it becomes the very thing that is unknown and is investigated at the greatest remove. Sometimes knowledge that is close at hand is like the reminder, and despite the least amount of effort it was like something overlooked, and so awareness does not turn to pursue it because it weakly understands it, in which case one needs to take a remote position in relation to it. From [all of] this, it has become clear that these faculties have a gathering place to which all of them can be traced back, and that it is not a body, regardless of whether it is or is not joined with the body.

NOTES

1. If it is the latter, then the organ would be just part, albeit an essential one, of what is identified as the self.
2. That is, what would be doing the perceiving?

On the Hypothesis that Animals Are Automata, and Its History

Thomas H. Huxley

. . . Thus far, the propositions respecting the physiology of the nervous system which are stated by Descartes have simply been more clearly defined, more fully illustrated, and, for the most part, demonstrated, by modern physiological research. But there remains a doctrine to which Descartes attached great weight, so that full acceptance of it became a sort of note of a thoroughgoing Cartesian, but which, nevertheless, is so opposed to ordinary prepossessions that it attained more general notoriety, and gave rise to more discussion, than almost any other Cartesian hypothesis. It is the doctrine that brute animals are mere machines or automata, devoid not only of reason, but of any kind of consciousness, which is stated briefly in the 'Discours de la Méthode,' and more fully in the 'Réponses aux Quatrièmes Objections,' and in the correspondence with Henry More.[1]

The process of reasoning by which Descartes arrived at this startling conclusion is well shown in the following passage of the 'Réponses':—

> But as regards the souls of beasts, although this is not the place for considering them, and though, without a general exposition of physics, I can say no more on this subject than I have already said in the fifth part of my Treatise on Method; yet, I will further state, here, that it appears to me to be a very remarkable circumstance that no movement can take place, either in the bodies of beasts, or even in our own, if these bodies have not in themselves all the organs and instruments by means of which the very same movements would be accomplished in a machine. So that, even in us, the spirit, or the soul, does not directly move the limbs, but only determines the course of that very subtle liquid which is called the animal spirits, which, running continually from the heart by the brain into the muscles, is the cause of all the movements of our limbs, and often may cause many different motions, one as easily as the other.
>
> And it does not even always exert this determination; for among the movements which take place in us, there are many which do not depend on the mind at all, such as the beating of the heart, the digestion of food, the nutrition, the respiration of those who sleep; and even in those who are awake, walking, singing, and other similar actions, when they are performed without the mind thinking about them. And, when one who falls from a height throws his hands forward to save his head, it is in virtue of

Excerpted from *Fortnightly Review* 16 (1874): pp. 555–80.

no ratiocination that he performs this action; it does not depend upon his mind, but takes place merely because his senses being affected by the present danger, some change arises in his brain which determines the animal spirits to pass thence into the nerves, in such a manner as is required to produce this motion, in the same way as in a machine, and without the mind being able to hinder it. Now since we observe this in ourselves, why should we be so much astonished if the light reflected from the body of a wolf into the eye of a sheep has the same force to excite in it the motion of flight?

After having observed this, if we wish to learn by reasoning, whether certain movements of beasts are comparable to those which are effected in us by the operation of the mind, or, on the contrary, to those which depend only on the animal spirits and the disposition of the organs, it is necessary to consider the difference between the two, which I have explained in the fifth part of the Discourse on Method (for I do not think that any others are discoverable), and then it will easily be seen, that all the actions of beasts are similar only to those which we perform without the help of our minds. For which reason we shall be forced to conclude, that we know of the existence in them of no other principle of motion than the disposition of their organs and the continual affluence of animal spirits produced by the heat of the heart, which attenuates and subtilises the blood; and, at the same time, we shall acknowledge that we have had no reason for assuming any other principle, except that, not having distinguished these two principles of motion, and seeing that the one, which depends only on the animal spirits and the organs, exists in beasts as well as in us, we have hastily concluded that the other, which depends on mind and on thought, was also possessed by them.

Descartes' line of argument is perfectly clear. He starts from reflex action in man, from the unquestionable fact that, in ourselves, co-ordinate, purposive, actions may take place, without the intervention of consciousness or volition, or even contrary to the latter. As actions of a certain degree of complexity are brought about by mere mechanism, why may not actions of still greater complexity be the result of a more refined mechanism? What proof is there that brutes are other than a superior race of marionettes, which eat without pleasure, cry without pain, desire nothing, know nothing, and only simulate intelligence as a bee simulates a mathematician?[2]

The Port Royalists adopted the hypothesis that brutes are machines, and are said to have carried its practical applications so far as to treat domestic animals with neglect, if not with actual cruelty. As late as the middle of the eighteenth century, the problem was discussed very fully and ably by Bouillier, in his 'Essai philosophique sur l'Ame des Bêtes,' while Condillac deals with it in his 'Traite des Animaux'; but since then it has received little attention. Nevertheless, modern research has brought to light a great multitude of facts, which not only show that Descartes' view is defensible, but render it far more defensible than it was in his day.

It must be premised, that it is wholly impossible absolutely to prove the presence or absence of consciousness in anything but one's own brain, though, by analogy, we are justified in assuming its existence in other men. Now if, by some accident, a man's spinal cord is divided, his limbs are paralysed, so far as his volition is concerned, below the point of injury; and he is incapable of experiencing all those states of consciousness which, in his uninjured state, would be excited by irritation of those nerves which come off below the injury. If the spinal cord is divided in the middle of the back, for example, the skin of the feet may be cut, or pinched, or burned, or wetted with vitriol, without any sensation of touch, or of pain, arising in consciousness. So far as the man is concerned, therefore, the part of the central nervous system which lies beyond the injury is cut off from consciousness. It must indeed be admitted, that, if any one think fit to maintain that the spinal cord below the injury is conscious, but that it is cut off from any means of making its consciousness known to the other consciousness in the brain, there is no means of driving him from his position by logic. But assuredly there is no way of proving it, and in the matter of consciousness, if in anything, we may hold by the rule, 'De non apparentibus et de non existentibus eadem est ratio.' However near the brain the spinal cord is injured, consciousness remains intact, except that the irritation of parts below the injury is no longer represented by sensation. On the other hand, pressure upon the anterior division of the brain, or extensive injuries to it, abolish consciousness. Hence, it is a highly probable conclusion, that consciousness in man depends upon the integrity of the anterior division of the brain, while the middle and hinder divisions of the brain,[3] and the rest of the nervous centres, have nothing to do with it. And it is further highly probable, that what is true for man is true for other vertebrated animals.

We may assume, then, that in a living vertebrated animal, any segment of the cerebro-spinal axis (or spinal cord and brain) separated from that anterior division of the brain which is the organ of consciousness, is as completely incapable of giving rise to consciousness as we know it to be incapable of carrying out volitions. Nevertheless, this separated segment of the spinal cord is not passive and inert. On the contrary, it is the seat of extremely remarkable powers. In our imaginary case of injury, the man would, as we have seen, be devoid of sensation in his legs, and would have not the least power of moving them. But, if the soles of his feet were tickled, the legs would be drawn up just as vigorously as they would have been before the injury. We know exactly what happens when the soles of the feet are tickled; a molecular change takes place in the sensory nerves of the skin, and is propagated along them and through the posterior roots of the spinal nerves, which are constituted by them, to the grey matter of the spinal cord. Through that grey matter the molecular motion is reflected into the anterior roots of the same nerves, constituted by the filaments which supply the muscles of the legs, and, travelling along these motor filaments, reaches the muscles, which at once contract, and cause the limbs to be drawn up.

In order to move the legs in this way, a definite co-ordination of muscular contractions is necessary; the muscles must contract in a certain order and with duly proportioned force; and moreover, as the feet are drawn away from the source of irritation, it may be said that the action has a final cause, or is purposive.

Thus it follows, that the grey matter of the segment of the man's spinal cord, though it is devoid of consciousness, nevertheless responds to a simple stimulus by giving rise to a complex set of muscular contractions, co-ordinated towards a definite end, and serving an obvious purpose.

If the spinal cord of a frog is cut across, so as to provide us with a segment separated from the brain, we shall have a subject parallel to the injured man, on which experiments can be made without remorse; as we have a right to conclude that a frog's spinal cord is not likely to be conscious, when a man's is not.

Now the frog behaves just as the man did. The legs are utterly paralysed, so far as voluntary movement is concerned; but they are vigorously drawn up to the body when any irritant is applied to the foot. But let us study our frog a little farther. Touch the skin of the side of the body with a little acetic acid, which gives rise to all the signs of great pain in an uninjured frog. In this case, there can be no pain, because the application is made to a part of the skin supplied with nerves which come off from the cord below the point of section; nevertheless, the frog lifts up the limb of the same side, and applies the foot to rub off the acetic acid; and, what is still more remarkable, if the limb be held so that the frog cannot use it, it will, by and by, move the limb of the other side, turn it across the body, and use it for the same rubbing process. It is impossible that the frog, if it were in its entirety and could reason, should perform actions more purposive than these: and yet we have most complete assurance that, in this case, the frog is not acting from purpose, has no consciousness, and is a mere insensible machine.

But now suppose that, instead of making a section of the cord in the middle of the body, it had been made in such a manner as to separate the hindermost division of the brain from the rest of the organ, and suppose the foremost two-thirds of the brain entirely taken away. The frog is then absolutely devoid of any spontaneity; it sits upright in the attitude which a frog habitually assumes; and it will not stir unless it is touched; but it differs from the frog which I have just described in this, that, if it be thrown into the water, it begins to swim, and swims just as well as the perfect frog does. But swimming requires the combination and successive co-ordination of a great number of muscular actions. And we are forced to conclude, that the impression made upon the sensory nerves of the skin of the frog by the contact with the water into which it is thrown, causes the transmission to the central nervous apparatus of an impulse which sets going a certain machinery by which all the muscles of swimming are brought into play in due co-ordination. If the frog be stimulated by some irritating body, it jumps or walks as well as the complete frog can do. The simple sensory impression, acting through the machinery of the cord, gives rise to these complex combined movements.

It is possible to go a step farther. Suppose that only the anterior division of the brain—so much of it as lies in front of the 'optic lobes'—is removed. If that operation is performed quickly and skilfully, the frog may be kept in a state of full bodily vigour for months, or it may be for years; but it will sit unmoved. It sees nothing: it hears nothing. It will starve sooner than feed itself, although food put into its mouth is swallowed. On irritation, it jumps or walks; if

thrown into the water it swims. If it be put on the hand, it sits there, crouched, perfectly quiet, and would sit there for ever. If the hand be inclined very gently and slowly, so that the frog would naturally tend to slip off, the creature's fore paws are shifted on to the edge of the hand, until he can just prevent himself from falling. If the turning of the hand be slowly continued, he mounts up with great care and deliberation, putting first one leg forward and then another, until he balances himself with perfect precision upon the edge; and if the turning of the hand is continued, he goes through the needful set of muscular operations, until he comes to be seated in security, upon the back of the hand. The doing of all this requires a delicacy of co-ordination, and a precision of adjustment of the muscular apparatus of the body, which are only comparable to those of a rope-dancer. To the ordinary influences of light, the frog, deprived of its cerebral hemispheres, appears to be blind. Nevertheless, if the animal be put upon a table, with a book at some little distance between it and the light, and the skin of the hinder part of its body is then irritated, it will jump forward, avoiding the book by passing to the right or left of it. Therefore, although the frog appears to have no sensation of light, visible objects act through its brain upon the motor mechanism of its body.[4]

It is obvious, that had Descartes been acquainted with these remarkable results of modern research, they would have furnished him with far more powerful arguments than he possessed in favour of his view of the automatism of brutes. The habits of a frog, leading its natural life, involve such simple adaptations to surrounding conditions, that the machinery which is competent to do so much without the intervention of consciousness, might well do all. And this argument is vastly strengthened by what has been learned in recent times of the marvellously complex operations which are performed mechanically, and to all appearance without consciousness, by men, when, in consequence of injury or disease, they are reduced to a condition more or less comparable to that of a frog, in which the anterior part of the brain has been removed. A case has recently been published by an eminent French physician, Dr. Mesnet, which illustrates this condition so remarkably, that I make no apology for dwelling upon it at considerable length.[5]

A sergeant of the French army, F—, twenty-seven years of age, was wounded during the battle of Bazeilles, by a ball which fractured his left parietal bone. He ran his bayonet through the Prussian soldier who wounded him, but almost immediately his right arm became paralysed; after walking about two hundred yards, his right leg became similarly affected, and he lost his senses. When he recovered them, three weeks afterwards, in hospital at Mayence, the right half of the body was completely paralysed, and remained in this condition for a year. At present, the only trace of the paralysis which remains is a slight weakness of the right half of the body. Three or four months after the wound was inflicted, periodical disturbances of the functions of the brain made their appearance, and have continued ever since. The disturbances last from fifteen to thirty hours; the intervals at which they occur being from fifteen to thirty days.

For four years, therefore, the life of this man has been divided into alternating phases—short abnormal states intervening between long normal states.

In the periods of normal life, the ex-sergeant's health is perfect; he is intelligent and kindly, and performs, satisfactorily, the duties of a hospital attendant. The commencement of the abnormal state is ushered in by uneasiness and a sense of weight about the forehead, which the patient compares to the constriction of a circle of iron; and, after its termination, he complains, for some hours, of dulness and heaviness of the head. But the transition from the normal to the abnormal state takes place in a few minutes, without convulsions or cries, and without anything to indicate the change to a bystander. His movements remain free and his expression calm, except for a contraction of the brow, an incessant movement of the eyeballs, and a chewing motion of the jaws. The eyes are wide open, and their pupils dilated. If the man happens to be in a place to which he is accustomed, he walks about as usual; but, if he is in a new place, or if obstacles are intentionally placed in his way, he stumbles gently against them, stops, and then, feeling over the objects with his hands, passes on one side of them. He offers no resistance to any change of direction which may be impressed upon him, or to the forcible acceleration or retardation of his movements. He eats, drinks, smokes, walks about, dresses and undresses himself, rises and goes to bed at the accustomed hours. Nevertheless, pins may be run into his body, or strong electric shocks sent through it, without causing the least indication of pain; no odorous substance, pleasant or unpleasant, makes the least impression; he

eats and drinks with avidity whatever is offered, and takes asafœtida, or vinegar, or quinine, as readily as water; no noise affects him; and light influences him only under certain conditions. Dr. Mesnet remarks, that the sense of touch alone seems to persist, and indeed to be more acute and delicate than in the normal state: and it is by means of the nerves of touch, almost exclusively, that his organism is brought into relation with the external world. Here a difficulty arises. It is clear from the facts detailed, that the nervous apparatus by which, in the normal state, sensations of touch are excited, is that by which external influences determine the movements of the body, in the abnormal state. But does the state of consciousness, which we term a tactile sensation, accompany the operation of this nervous apparatus in the abnormal state? or is consciousness utterly absent, the man being reduced to an insensible mechanism? . . .

As I have pointed out, it is impossible to prove that F— is absolutely unconscious in his abnormal state, but it is no less impossible to prove the contrary; and the case of the frog goes a long way to justify the assumption that, in the abnormal state, the man is a mere insensible machine.

If such facts as these had come under the knowledge of Descartes, would they not have formed an apt commentary upon that remarkable passage in the 'Traité de l'Homme,' which I have quoted elsewhere, but which is worth repetition?—

> All the functions which I have attributed to this machine (the body), as the digestion of food, the pulsation of the heart and of the arteries; the nutrition and the growth of the limbs; respiration, wakefulness, and sleep; the reception of light, sounds, odours, flavours, heat, and such like qualities, in the organs of the external senses; the impression of the ideas of these in the organ of common sensation and in the imagination; the retention or the impression of these ideas on the memory; the internal movements of the appetites and the passions; and lastly the external movements of all the limbs, which follow so aptly, as well the action of the objects which are presented to the senses, as the impressions which meet in the memory, that they imitate as nearly as possible those of a real man; I desire, I say, that you should consider that these functions in the machine naturally proceed from the mere arrangement of its organs, neither more nor less than do the movements of a clock, or other automaton, from that of its weights and its wheels; so that, so far as these are concerned, it is not necessary to conceive any other vegetative or sensitive soul, nor any other principle of motion or of life, than the blood and the spirits agitated by the fire which burns continually in the heart, and which is no wise essentially different from all the fires which exist in inanimate bodies.

And would Descartes not have been justified in asking why we need deny that animals are machines, when men, in a state of unconsciousness, perform, mechanically, actions as complicated and as seemingly rational as those of any animals?

But though I do not think that Descartes' hypothesis can be positively refuted, I am not disposed to accept it. The doctrine of continuity is too well established for it to be permissible to me to suppose that any complex natural phenomenon comes into existence suddenly, and without being preceded by simpler modifications; and very strong arguments would be needed to prove that such complex phenomena as those of consciousness, first make their appearance in man. We know, that, in the individual man, consciousness grows from a dim glimmer to its full light, whether we consider the infant advancing in years, or the adult emerging from slumber and swoon. We know, further, that the lower animals possess, though less developed, that part of the brain which we have every reason to believe to be the organ of consciousness in man; and as, in other cases, function and organ are proportional, so we have a right to conclude it is with the brain; and that the brutes, though they may not possess our intensity of consciousness, and though, from the absence of language, they can have no trains of thoughts, but only trains of feelings, yet have a consciousness which, more or less distinctly, foreshadows our own.

I confess that, in view of the struggle for existence which goes on in the animal world, and of the frightful quantity of pain with which it must be accompanied, I should be glad if the probabilities were in favour of Descartes' hypothesis; but, on the other hand, considering the terrible practical consequences to domestic animals which might ensue from any error on our part, it is as well to err on the right side, if we err at all, and deal with them as weaker brethren, who are bound, like the rest of us, to pay their toll for living, and suffer what is needful for the general good. As Hartley finely says, 'We seem to be in the place of God to them'; and we may justly follow the precedents He sets in nature in our dealings with them.

But though we may see reason to disagree with Descartes' hypothesis that brutes are

unconscious machines, it does not follow that he was wrong in regarding them as automata. They may be more or less conscious, sensitive, automata; and the view that they are such conscious machines is that which is implicitly, or explicitly, adopted by most persons. When we speak of the actions of the lower animals being guided by instinct and not by reason, what we really mean is that, though they feel as we do, yet their actions are the results of their physical organisation. We believe, in short, that they are machines, one part of which (the nervous system) not only sets the rest in motion, and co-ordinates its movements in relation with changes in surrounding bodies, but is provided with special apparatus, the function of which is the calling into existence of those states of consciousness which are termed sensations, emotions, and ideas. I believe that this generally accepted view is the best expression of the facts at present known.

It is experimentally demonstrable—any one who cares to run a pin into himself may perform a sufficient demonstration of the fact—that a mode of motion of the nervous system is the immediate antecedent of a state of consciousness. All but the adherents of 'Occasionalism,' or of the doctrine of 'Pre-established Harmony' (if any such now exist), must admit that we have as much reason for regarding the mode of motion of the nervous system as the cause of the state of consciousness, as we have for regarding any event as the cause of another. How the one phenomenon causes the other we know, as much or as little, as in any other case of causation; but we have as much right to believe that the sensation is an effect of the molecular change, as we have to believe that motion is an effect of impact; and there is as much propriety in saying that the brain evolves sensation, as there is in saying that an iron rod, when hammered, evolves heat.

As I have endeavoured to show, we are justified in supposing that something analogous to what happens in ourselves takes place in the brutes, and that the affections of their sensory nerves give rise to molecular changes in the brain, which again give rise to, or evolve, the corresponding states of consciousness. Nor can there be any reasonable doubt that the emotions of brutes, and such ideas as they possess, are similarly dependent upon molecular brain changes. Each sensory impression leaves behind a record in the structure of the brain—an 'ideagenous' molecule, so to speak, which is competent, under certain conditions, to reproduce, in a fainter condition, the state of consciousness which corresponds with that sensory impression; and it is these 'ideagenous molecules' which are the physical basis of memory.

It may be assumed, then, that molecular changes in the brain are the causes of all the states of consciousness of brutes. Is there any evidence that these states of consciousness may, conversely, cause those molecular changes which give rise to muscular motion? I see no such evidence. The frog walks, hops, swims, and goes through his gymnastic performances quite as well without consciousness, and consequently without volition, as with it; and, if a frog, in his natural state, possesses anything corresponding with what we call volition, there is no reason to think that it is anything but a concomitant of the molecular changes in the brain which form part of the series involved in the production of motion.

The consciousness of brutes would appear to be related to the mechanism of their body simply as a collateral product of its working, and to be as completely without any power of modifying that working as the steam-whistle which accompanies the work of a locomotive engine is without influence upon its machinery. Their volition, if they have any, is an emotion indicative of physical changes, not a cause of such changes.

This conception of the relations of states of consciousness with molecular changes in the brain—of *psychoses* with *neuroses*—does not prevent us from ascribing free will to brutes. For an agent is free when there is nothing to prevent him from doing that which he desires to do. If a greyhound chases a hare, he is a free agent, because his action is in entire accordance with his strong desire to catch the hare; while so long as he is held back by the leash he is not free, being prevented by external force from following his inclination. And the ascription of freedom to the greyhound under the former circumstances is by no means inconsistent with the other aspect of the facts of the case—that he is a machine impelled to the chase, and caused, at the same time, to have the desire to catch the game by the impression which the rays of light proceeding from the hare make upon his eyes, and through them upon his brain.

Much ingenious argument has at various times been bestowed upon the question: How is it possible to imagine that volition, which is a state of consciousness, and, as such, has not the slightest community of nature with matter

in motion, can act upon the moving matter of which the body is composed, as it is assumed to do in voluntary acts? But if, as is here suggested, the voluntary acts of brutes—or, in other words, the acts which they desire to perform—are as purely mechanical as the rest of their actions, and are simply accompanied by the state of consciousness called volition, the inquiry, so far as they are concerned, becomes superfluous. Their volitions do not enter into the chain of causation of their actions at all.

The hypothesis that brutes are conscious automata is perfectly consistent with any view that may be held respecting the often discussed and curious question whether they have souls or not; and, if they have souls, whether those souls are immortal or not. It is obviously harmonious with the most literal adherence to the text of Scripture concerning 'the beast that perisheth'; but it is not inconsistent with the amiable conviction ascribed by Pope to his 'untutored savage,' that when he passes to the happy hunting-grounds in the sky, 'his faithful dog shall bear him company.' If the brutes have consciousness and no souls, then it is clear that, in them, consciousness is a direct function of material changes; while, if they possess immaterial subjects of consciousness, or souls, then, as consciousness is brought into existence only as the consequence of molecular motion of the brain, it follows that it is an indirect product of material changes. The soul stands related to the body as the bell of a clock to the works, and consciousness answers to the sound which the bell gives out when it is struck.

Thus far I have strictly confined myself to the problem with which I proposed to deal at starting—the automatism of brutes. The question is, I believe, a perfectly open one, and I feel happy in running no risk of either Papal or Presbyterian condemnation for the views which I have ventured to put forward. And there are so very few interesting questions which one is, at present, allowed to think out scientifically—to go as far as reason leads, and stop where evidence comes to an end—without speedily being deafened by the tattoo of 'the drum ecclesiastic'—that I have luxuriated in my rare freedom, and would now willingly bring this disquisition to an end if I could hope that other people would go no farther. Unfortunately, past experience debars me from entertaining any such hope, even if

> *. . . that drum's discordant sound Parading round and round and round,*

were not, at present, as audible to me as it was to the mild poet who ventured to express his hatred of drums in general, in that well-known couplet.

It will be said, that I mean that the conclusions deduced from the study of the brutes are applicable to man, and that the logical consequences of such application are fatalism, materialism, and atheism—whereupon the drums will beat the *pas de charge.*

One does not do battle with drummers; but I venture to offer a few remarks for the calm consideration of thoughtful persons, untrammelled by foregone conclusions, unpledged to shore-up tottering dogmas, and anxious only to know the true bearings of the case.

It is quite true that, to the best of my judgment, the argumentation which applies to brutes holds equally good of men; and, therefore, that all states of consciousness in us, as in them, are immediately caused by molecular changes of the brain-substance. It seems to me that in men, as in brutes, there is no proof that any state of consciousness is the cause of change in the motion of the matter of the organism. If these positions are well based, it follows that our mental conditions are simply the symbols in consciousness of the changes which takes place automatically in the organism; and that, to take an extreme illustration, the feeling we call volition is not the cause of a voluntary act, but the symbol of that state of the brain which is the immediate cause of that act. We are conscious automata, endowed with free will in the only intelligible sense of that much-abused term—inasmuch as in many respects we are able to do as we like—but nonetheless parts of the great series of causes and effects which, in unbroken continuity, composes that which is, and has been, and shall be—the sum of existence.

NOTES

1. *Réponse de M. Descartes a M. Morus.* 1649. *Œuevres,* tome x. p. 204. 'Mais le plus grand de tous les préjugés que nous ayons retenus de notre enfance, est celui de croire que les bêtes pensent,' etc.

2. Malebranche states the view taken by orthodox Cartesians in 1689 very forcibly: 'Ainsi dans les chiens, les chats, et les autres animaux, il n'y a ny intelligence, ny âme spirituelle comme on l'entend ordinairement. Ils mangent sans plaisir;

ils crient sans douleur; ils croissent sans le sçavoir; ils ne desirent rien; ils ne connoissent rien; et s'ils agissent avec adresse et d'une maniere qui marque l'intelligence, c'est que Dieu les faisant pour les conserver, il a conformé leurs corps de telle manière, qu'ils évitent organiquement, sans le sçavoir, tout ce qui peut les de truire et qu'ils semblent craindre.' *Feuillet de Conches. Méditations Métaphysiques et Correspondance de. N. Malebranche. Neuvième Méditation.* 1841.

3. Not to be confounded with the anterior middle and hinder parts of the hemispheres of the cerebrum.
4. See the remarkable essay of Göltz, *Beitrage zur Lehre von den Functionen der Nervencentren des Frosches*, published in 1809. I have repeated Göltz's experiments, and obtained the same results.
5. "De l'Automatisme de la Mémoire et du Souvenir, dans le Somnambulisme pathologique." Par le Dr. E. Mesnet, Médecin de l'Hôpital Saint-Antoine. *L'Union Médicale*, Juillet 21 et 23, 1874. My attention was first called to a summary of this remarkable case, which appeared in the *Journal des Débats* for the 7th of August, 1874, by my friend General Strachey, F. R. S.

An Unfortunate Dualist

Raymond M. Smullyan

Once upon a time there was a dualist. He believed that mind and matter are separate substances. Just how they interacted he did not pretend to know—this was one of the 'mysteries' of life. But he was sure they were quite separate substances.

This dualist, unfortunately, led an unbearably painful life—not because of his philosophical beliefs, but for quite different reasons. And he had excellent empirical evidence that no respite was in sight for the rest of his life. He longed for nothing more than to die. But he was deterred from suicide by such reasons as: (1) he did not want to hurt other people by his death; (2) he was afraid suicide might be morally wrong; (3) he was afraid there *might* be an afterlife, and he did not want to risk the possibility of eternal punishment. So our poor dualist was quite desperate.

Then came the discovery of *the* miracle drug! Its effect on the taker was to annihilate the soul or mind entirely but to leave the body functioning *exactly* as before. Absolutely no observable change came over the taker; the body continued to act just as if it still had a soul. Not the closest friend or observer could possibly know that the taker had taken the drug, unless the taker informed him.

Do you believe that such a drug is impossible in principle? Assuming you believe it possible, would you take it? Would you regard it as immoral? Is it tantamount to suicide? Is there anything in Scriptures forbidding the use of such a drug? Surely, the *body* of the taker can still fulfill all its responsibilities on earth. Another question: Suppose your spouse took such a drug, and you knew it. You would know that she (or he) no longer had a soul but acted just as if she did have one. Would you love your mate any less?

To return to the story, our dualist was, of course, delighted! Now he could annihilate himself (his *soul*, that is) in a way not subject to any of the foregoing objections. And so, for the first time in years, he went to bed with a light heart, saying: 'Tomorrow morning I will go down to the drugstore and get the drug. My days of suffering are over at last!' With these thoughts, he fell peacefully asleep.

Now at this point a curious thing happened. A friend of the dualist who knew about this drug, and who knew of the sufferings of the dualist, decided to put him out of his misery. So in the middle of the night, while the dualist was fast asleep, the friend quietly stole into the house and injected the drug into his veins. The next morning the body of the dualist awoke—without any soul indeed—and the first thing it did was to go to the drugstore to get the drug. He took it home and, before taking it, said, 'Now I shall be released.' So he took it and then waited the time interval in which it was supposed to work. At the end of the interval he angrily exclaimed: 'Damn it, this stuff hasn't helped at all! I still obviously have a soul and am suffering as much as ever!'

Doesn't all this suggest that perhaps there might be something just a *little* wrong with dualism?

Excerpted from Raymond Smullyan, *This Book Needs No Title* (Upper Saddle River, NJ: Prentice-Hall, 1980), pp. 53–55. Reprinted by permission of the author.

Descartes' Myth

Gilbert Ryle

1. The Official Doctrine

There is a doctrine about the nature and place of minds which is so prevalent among theorists and even among laymen that it deserves to be described as the official theory. Most philosophers, psychologists and religious teachers subscribe, with minor reservations, to its main articles and, although they admit certain theoretical difficulties in it, they tend to assume that these can be overcome without serious modifications being made to the architecture of the theory. It will be argued here that the central principles of the doctrine are unsound and conflict with the whole body of what we know about minds when we are not speculating about them.

The official doctrine, which hails chiefly from Descartes, is something like this. With the doubtful exceptions of idiots and infants in arms every human being has both a body and a mind. Some would prefer to say that every human being is both a body and a mind. His body and his mind are ordinarily harnessed together, but after the death of the body his mind may continue to exist and function.

Human bodies are in space and are subject to the mechanical laws which govern all other bodies in space. Bodily processes and states can be inspected by external observers. So a man's bodily life is as much a public affair as are the lives of animals and reptiles and even as the careers of trees, crystals and planets.

But minds are not in space, nor are their operations subject to mechanical laws. The workings of one mind are not witnessable by other observers; its career is private. Only I can take direct cognisance of the states and processes of my own mind. A person therefore lives through two collateral histories, one consisting of what happens in and to his body, the other consisting of what happens in and to his mind. The first is public, the second private. The events in the first history are events in the physical world, those in the second are events in the mental world.

It has been disputed whether a person does or can directly monitor all or only some of the episodes of his own private history; but, according to the official doctrine, of at least some of these episodes he has direct and unchallengeable cognisance. In consciousness, self-consciousness and introspection he is directly and authentically apprised of the present states and operations of his mind. He may have great or small uncertainties about concurrent and adjacent episodes in the physical world, but he can have none about at least part of what is momentarily occupying his mind.

It is customary to express this bifurcation of his two lives and of his two worlds by saying that the things and events which belong to the physical world, including his own body, are external, while the workings of his own mind are internal. This antithesis of outer and inner is of course meant to be construed as a metaphor, since minds, not being in space, could not be described as being spatially inside anything else, or as having things going on spatially inside themselves. But relapses from this good intention are common and theorists are found speculating how stimuli, the physical sources of which are yards or miles outside a person's skin, can generate mental responses inside his skull, or how decisions framed inside his cranium can set going movements of his extremities.

Even when 'inner' and 'outer' are construed as metaphors, the problem how a person's mind

From Gilbert Ryle, *The Concept of Mind* (Hutchinson, 1949), pp. 11–24. Reprinted with permission of the publisher and the Principal, Fellows, and Scholars of Hertford College in the University of Oxford.

and body influence one another is notoriously charged with theoretical difficulties. What the mind wills, the legs, arms and the tongue execute; what affects the ear and the eye has something to do with what the mind perceives; grimaces and smiles betray the mind's moods and bodily castigations lead, it is hoped, to moral improvement. But the actual transactions between the episodes of the private history and those of the public history remain mysterious, since by definition they can belong to neither series. They could not be reported among the happenings described in a person's autobiography of his inner life, but nor could they be reported among those described in some one else's biography of that person's overt career. They can be inspected neither by introspection nor by laboratory experiment. They are theoretical shuttle-cocks which are forever being bandied from the physiologist back to the psychologist and from the psychologist back to the physiologist.

Underlying this partly metaphorical representation of the bifurcation of a person's two lives there is a seemingly more profound and philosophical assumption. It is assumed that there are two different kinds of existence or status. What exists or happens may have the status of physical existence, or it may have the status of mental existence. Somewhat as the faces of coins are either heads or tails, or somewhat as living creatures are either male or female, so, it is supposed, some existing is physical existing, other existing is mental existing. It is a necessary feature of what has physical existence that it is in space and time; it is a necessary feature of what has mental existence that it is in time but not in space. What has physical existence is composed of matter, or else is a function of matter; what has mental existence consists of consciousness, or else is a function of consciousness.

There is thus a polar opposition between mind and matter, an opposition which is often brought out as follows. Material objects are situated in a common field, known as 'space,' and what happens to one body in one part of space is mechanically connected with what happens to other bodies in other parts of space. But mental happenings occur in insulated fields, known as 'minds,' and there is, apart maybe from telepathy, no direct causal connection between what happens in one mind and what happens in another. Only through the medium of the public physical world can the mind of one person make a difference to the mind of another. The mind is its own place and in his inner life each of us lives the life of a ghostly Robinson Crusoe. People can see, hear and jolt one another's bodies, but they are irremediably blind and deaf to the workings of one another's minds and inoperative upon them.

What sort of knowledge can be secured of the workings of a mind? On the one side, according to the official theory, a person has direct knowledge of the best imaginable kind of the workings of his own mind. Mental states and processes are (or are normally) conscious states and processes, and the consciousness which irradiates them can engender no illusions and leaves the door open for no doubts. A person's present thinkings, feelings and willings, his perceivings, rememberings and imaginings are intrinsically 'phos-phorescent', their existence and their nature are inevitably betrayed to their owner. The inner life is a stream of consciousness of such a sort that it would be absurd to suggest that the mind whose life is that stream might be unaware of what is passing down it.

True, the evidence adduced recently by Freud seems to show that there exist channels tributary to this stream, which run hidden from their owner. People are actuated by impulses the existence of which they vigorously disavow; some of their thoughts differ from the thoughts which they acknowledge; and some of the actions which they think they will to perform they do not really will. They are thoroughly gulled by some of their own hypocrisies and they successfully ignore facts about their mental lives which on the official theory ought to be patent to them. Holders of the official theory tend, however, to maintain that anyhow in normal circumstances a person must be directly and authentically seized of the present state and workings of his own mind.

Besides being currently supplied with these alleged immediate data of consciousness, a person is also generally supposed to be able to exercise from time to time a special kind of perception, namely inner perception, or introspection. He can take a (non-optical) 'look' at what is passing in his mind. Not only can he view and scrutinize a flower through his sense of sight and listen to and discriminate the notes of a bell through his sense of hearing; he can also reflectively or introspectively watch, without any bodily organ of sense, the current episodes of his inner life. This self-observation is also commonly supposed to be immune from illusion, confusion or doubt. A mind's reports of its own affairs have a certainty superior to the

best that is possessed by its reports of matters in the physical world. Sense-perceptions can, but consciousness and introspection cannot, be mistaken or confused.

On the other side, one person has no direct access of any sort to the events of the inner life of another. He cannot do better than make problematic inferences from the observed behaviour of the other person's body to the states of mind which, by analogy from his own conduct, he supposes to be signalised by that behaviour. Direct access to the workings of a mind is the privilege of that mind itself; in default of such privileged access, the workings of one mind are inevitably occult to everyone else. For the supposed arguments from bodily movements similar to their own to mental workings similar to their own would lack any possibility of observational corroboration. Not unnaturally, therefore, an adherent of the official theory finds it difficult to resist this consequence of his premisses, that he has no good reason to believe that there do exist minds other than his own. Even if he prefers to believe that to other human bodies there are harnessed minds not unlike his own, he cannot claim to be able to discover their individual characteristics, or the particular things that they undergo and do. Absolute solitude is on this showing the ineluctable destiny of the soul. Only our bodies can meet.

As a necessary corollary of this general scheme there is implicitly prescribed a special way of construing our ordinary concepts of mental powers and operations. The verbs, nouns and adjectives, with which in ordinary life we describe the wits, characters and higher-grade performances of the people with whom we have to do, are required to be construed as signifying special episodes in their secret histories, or else as signifying tendencies for such episodes to occur. When someone is described as knowing, believing or guessing something, as hoping, dreading, intending or shirking something, as designing this or being amused at that, these verbs are supposed to denote the occurrence of specific modifications in his (to us) occult stream of consciousness. Only his own privileged access to this stream in direct awareness and introspection could provide authentic testimony that these mental-conduct verbs were correctly or incorrectly applied. The onlooker, be he teacher, critic, biographer or friend, can never assure himself that his comments have any vestige of truth. Yet it was just because we do in fact all know how to make such comments, make them with general correctness and correct them when they turn out to be confused or mistaken, that philosophers found it necessary to construct their theories of the nature and place of minds. Finding mental-conduct concepts being regularly and effectively used, they properly sought to fix their logical geography. But the logical geography officially recommended would entail that there could be no regular or effective use of these mental-conduct concepts in our descriptions of, and prescriptions for, other people's minds.

2. The Absurdity of the Official Doctrine

Such in outline is the official theory. I shall often speak of it, with deliberate abusiveness, as 'the dogma of the Ghost in the Machine.' I hope to prove that it is entirely false, and false not in detail but in principle. It is not merely an assemblage of particular mistakes. It is one big mistake and a mistake of a special kind. It is, namely, a category-mistake. It represents the facts of mental life as if they belonged to one logical type or category (or range of types or categories), when they actually belong to another. The dogma is therefore a philosopher's myth. In attempting to explode the myth I shall probably be taken to be denying well-known facts about the mental life of human beings, and my plea that I aim at doing nothing more than rectify the logic of mental-conduct concepts will probably be disallowed as mere subterfuge.

I must first indicate what is meant by the phrase 'Category-mistake.' This I do in a series of illustrations.

A foreigner visiting Oxford or Cambridge for the first time is shown a number of colleges, libraries, playing fields, museums, scientific departments and administrative offices. He then asks 'But where is the University? I have seen where the members of the Colleges live, where the Registrar works, where the scientists experiment and the rest. But I have not yet seen the University in which reside and work the members of your University.' It has then to be explained to him that the University is not another collateral institution, some ulterior counterpart to the colleges, laboratories and offices which he has seen. The University is just the way in which all that he has already seen is organized. When they are seen and when their coordination is understood, the University has been seen. His mistake lay in his innocent assumption that it was correct to speak of Christ Church,

the Bodleian Library, the Ashmolean Museum *and* the University, to speak, that is, as if 'the University' stood for an extra member of the class of which these other units are members. He was mistakenly allocating the University to the same category as that to which the other institutions belong.

The same mistake would be made by a child witnessing the march-past of a division, who, having had pointed out to him such and such battalions, batteries, squadrons, etc., asked when the division was going to appear. He would be supposing that a division was a counterpart to the units already seen, partly similar to them and partly unlike them. He would be shown his mistake by being told that in watching the battalions, batteries and squadrons marching past he had been watching the division marching past. The march-past was not a parade of battalions, batteries, squadrons *and* a division; it was a parade of the battalions, batteries and squadrons *of* a division.

One more illustration. A foreigner watching his first game of cricket learns what are the functions of the bowlers, the batsmen, the fielders, the umpires and the scorers. He then says 'But there is no one left on the field to contribute the famous element of team-spirit. I see who does the bowling, the batting and the wicket-keeping; but I do not see whose role it is to exercise *esprit de corps*.' Once more, it would have to be explained that he was looking for the wrong type of thing. Team-spirit is not another cricketing-operation supplementary to all of the other special tasks. It is, roughly, the keenness with which each of the special tasks is performed, and performing a task keenly is not performing two tasks. Certainly exhibiting team-spirit is not the same thing as bowling or catching, but nor is it a third thing such that we can say that the bowler first bowls *and* then exhibits team-spirit or that a fielder is at a given moment *either* catching *or* displaying *esprit de corps*.

These illustrations of category-mistakes have a common feature which must be noticed. The mistakes were made by people who did not know how to wield the concepts *University, division* and *team-spirit*. Their puzzles arose from inability to use certain items in the English vocabulary.

The theoretically interesting category-mistakes are those made by people who are perfectly competent to apply concepts, at least in the situations with which they are familiar, but are still liable in their abstract thinking to allocate those concepts to logical types to which they do not belong. An instance of a mistake of this sort would be the following story. A student of politics has learned the main differences between the British, the French and the American Constitutions, and has learned also the differences and connections between the Cabinet, Parliament, the various Ministries, the Judicature and the Church of England. But he still becomes embarrassed when asked questions about the connections between the Church of England, the Home Office and the British Constitution. For while the Church and the Home Office are institutions, the British Constitution is not another institution in the same sense of that noun. So inter-institutional relations which can be asserted or denied to hold between the Church and the Home Office cannot be asserted or denied to hold between either of them and the British Constitution. 'The British Constitution' is not a term of the same logical type as 'the Home Office' and 'the Church of England.' In a partially similar way, John Doe may be a relative, a friend, an enemy or a stranger to Richard Roe; but he cannot be any of these things to the Average Taxpayer. He knows how to talk sense in certain sorts of discussions about the Average Taxpayer, but he is baffled to say why he could not come across him in the street as he can come across Richard Roe.

It is pertinent to our main subject to notice that, so long as the student of politics continues to think of the British Constitution as a counterpart to the other institutions, he will tend to describe it as a mysteriously occult institution; and so long as John Doe continues to think of the Average Taxpayer as a fellow-citizen, he will tend to think of him as an elusive insubstantial man, a ghost who is everywhere yet nowhere.

My destructive purpose is to show that a family of radical category-mistakes is the source of the double-life theory. The representation of a person as a ghost mysteriously ensconced in a machine derives from this argument. Because, as is true, a person's thinking, feeling and purposive doing cannot be described solely in the idioms of physics, chemistry and physiology, therefore they must be described in counterpart idioms. As the human body is a complex organised unit, so the human mind must be another complex organised unit, though one made of a different sort of stuff and with a different sort of structure. Or, again, as the human body, like any other parcel of matter, is a field of causes

and effects, so the mind must be another field of causes and effects, though not (Heaven be praised) mechanical causes and effects.

3. The Origin of the Category-Mistake

One of the chief intellectual origins of what I have yet to prove to be the Cartesian category-mistake seems to be this. When Galileo showed that his methods of scientific discovery were competent to provide a mechanical theory which should cover every occupant of space, Descartes found in himself two conflicting motives. As a man of scientific genius he could not but endorse the claims of mechanics, yet as a religious and moral man he could not accept, as Hobbes accepted, the discouraging rider to those claims, namely that human nature differs only in degree of complexity from clockwork. The mental could not be just a variety of the mechanical.

He and subsequent philosophers naturally but erroneously availed themselves of the following escape-route. Since mental-conduct words are not to be construed as signifying the occurrence of mechanical processes, they must be construed as signifying the occurrence of non-mechanical processes; since mechanical laws explain movements in space as the effects of other movements in space, other laws must explain some of the non-spatial workings of minds as the effects of other non-spatial workings of minds. The difference between the human behaviours which we describe as intelligent and those which we describe as unintelligent must be a difference in their causation; so, while some movements of human tongues and limbs are the effects of mechanical causes, others must be the effects of non-mechanical causes, i.e. some issue from movements of particles of matter, others from workings of the mind.

The differences between the physical and the mental were thus represented as differences inside the common framework of the categories of 'thing,' 'stuff,' 'attribute,' 'state,' 'process,' 'change,' 'cause,' and 'effect.' Minds are things, but different sorts of things from bodies; mental processes are causes and effects, but different sorts of causes and effects from bodily movements. And so on. Somewhat as the foreigner expected the University to be an extra edifice, rather like a college but also considerably different, so the repudiators of mechanism represented minds as extra centres of causal processes, rather like machines but also considerably different from them. Their theory was a paramechanical hypothesis.

That this assumption was at the heart of the doctrine is shown by the fact that there was from the beginning felt to be a major theoretical difficulty in explaining how minds can influence and be influenced by bodies. How can a mental process, such as willing, cause spatial movements like the movements of the tongue? How can a physical change in the optic nerve have among its effects a mind's perception of a flash of light? This notorious crux by itself shows the logical mould into which Descartes pressed his theory of the mind. It was the self-same mould into which he and Galileo set their mechanics. Still unwittingly adhering to the grammar of mechanics, he tried to avert disaster by describing minds in what was merely an obverse vocabulary. The workings of minds had to be described by the mere negatives of the specific descriptions given to bodies; they are not in space, they are not motions, they are not modifications of matter, they are not accessible to public observation. Minds are not bits of clockwork, they are just bits of not-clockwork.

As thus represented, minds are not merely ghosts harnessed to machines, they are themselves just spectral machines. Though the human body is an engine, it is not quite an ordinary engine, since some of its workings are governed by another engine inside it—this interior governor-engine being one of a very special sort. It is invisible, inaudible and it has no size or weight. It cannot be taken to bits and the laws it obeys are not those known to ordinary engineers. Nothing is known of how it governs the bodily engine.

A second major crux points the same moral. Since, according to the doctrine, minds belong to the same category as bodies and since bodies are rigidly governed by mechanical laws, it seemed to many theorists to follow that minds must be similarly governed by rigid non-mechanical laws. The physical world is a deterministic system, so the mental world must be a deterministic system. Bodies cannot help the modifications that they undergo, so minds cannot help pursuing the careers fixed for them. *Responsibility, choice, merit* and *demerit* are therefore inapplicable concepts—unless the compromise solution is adopted of saying that the laws governing mental processes, unlike those governing physical processes, have the congenial attribute of being only rather rigid. The problem of the Freedom of the Will was the

problem how to reconcile the hypothesis that minds are to be described in terms drawn from the categories of mechanics with the knowledge that higher-grade human conduct is not of a piece with the behaviour of machines.

It is an historical curiosity that it was not noticed that the entire argument was broken-backed. Theorists correctly assumed that any sane man could already recognise the differences between, say, rational and non-rational utterances or between purposive and automatic behaviour. Else there would have been nothing requiring to be salved from mechanism. Yet the explanation given presupposed that one person could in principle never recognise the difference between the rational and the irrational utterances issuing from other human bodies, since he could never get access to the postulated immaterial causes of some of their utterances. Save for the doubtful exception of himself, he could never tell the difference between a man and a Robot. It would have to be conceded, for example, that, for all that we can tell, the inner lives of persons who are classed as idiots or lunatics are as rational as those of anyone else. Perhaps only their overt behaviour is disappointing; that is to say, perhaps 'idiots' are not really idiotic, or 'lunatics' lunatic. Perhaps, too, some of those who are classed as sane are really idiots. According to the theory, external observers could never know how the overt behaviour of others is correlated with their mental powers and processes and so they could never know or even plausibly conjecture whether their applications of mental-conduct concepts to these other people were correct or incorrect. It would then be hazardous or impossible for a man to claim sanity or logical consistency even for himself, since he would be debarred from comparing his own performances with those of others. In short, our characterisations of persons and their performances as intelligent, prudent and virtuous or as stupid, hypocritical and cowardly could never have been made, so the problem of providing a special causal hypothesis to serve as the basis of such diagnoses would never have arisen. The question, 'How do persons differ from machines?' arose just because everyone already knew how to apply mental-conduct concepts before the new causal hypothesis was introduced. This causal hypothesis could not therefore be the source of the criteria used in those applications. Nor, of course, has the causal hypothesis in any degree improved our handling of those criteria. We still distinguish good from bad arithmetic, politic from impolitic conduct and fertile from infertile imaginations in the ways in which Descartes himself distinguished them before and after he speculated how the applicability of these criteria was compatible with the principle of mechanical causation.

He had mistaken the logic of his problem. Instead of asking by what criteria intelligent behaviour is actually distinguished from non-intelligent behaviour, he asked, 'Given that the principle of mechanical causation does not tell us the difference, what other causal principle will tell it to us?' He realised that the problem was not one of mechanics and assumed that it must therefore be one of some counterpart to mechanics. Not unnaturally psychology is often cast for just this role.

When two terms belong to the same category, it is proper to construct conjunctive propositions embodying them. Thus a purchaser may say that he bought a left-hand glove and a right-hand glove, but not that he bought a left-hand glove, a right-hand glove and a pair of gloves. 'She came home in a flood of tears and a sedanchair' is a well-known joke based on the absurdity of conjoining terms of different types. It would have been equally ridiculous to construct the disjunction 'She came home either in a flood of tears or else in a sedan-chair.' Now the dogma of the Ghost in the Machine does just this. It maintains that there exist both bodies and minds; that there occur physical processes and mental processes; that there are mechanical causes of corporeal movements and mental causes of corporeal movements. I shall argue that these and other analogous conjunctions are absurd; but, it must be noticed, the argument will not show that either of the illegitimately conjoined propositions is absurd in itself. I am not, for example, denying that there occur mental processes. Doing long division is a mental process and so is making a joke. But I am saying that the phrase 'there occur mental processes' does not mean the same sort of thing as 'there occur physical processes,' and, therefore, that it makes no sense to conjoin or disjoin the two.

If my argument is successful, there will follow some interesting consequences. First, the hallowed contrast between Mind and Matter will be dissipated, but dissipated not by either of the equally hallowed absorptions of Mind by Matter or of Matter by Mind, but in quite a different way. For the seeming contrast of the two will be shown to be as illegitimate as would be the contrast of 'she came home in a flood

of tears' and 'she came home in a sedan-chair.' The belief that there is a polar opposition between Mind and Matter is the belief that they are terms of the same logical type.

It will also follow that both Idealism and Materialism are answers to an improper question. The 'reduction' of the material world to mental states and processes, as well as the 'reduction' of mental states and processes to physical states and processes, presuppose the legitimacy of the disjunction 'Either there exist minds or there exist bodies (but not both).' It would be like saying, 'Either she bought a left-hand and a right-hand glove or she bought a pair of gloves (but not both).'

It is perfectly proper to say, in one logical tone of voice, that there exist minds and to say, in another logical tone of voice, that there exist bodies. But these expressions do not indicate two different species of existence, for 'existence' is not a generic word like 'coloured' or 'sexed.' They indicate two different senses of 'exist,' somewhat as 'rising' has different senses in 'the tide is rising,' 'hopes are rising,' and 'the average age of death is rising.' A man would be thought to be making a poor joke who said that three things are now rising, namely the tide, hopes and the average age of death. It would be just as good or bad a joke to say that there exist prime numbers and Wednesdays and public opinions and navies; or that there exist both minds and bodies. In the succeeding chapters I try to prove that the official theory does rest on a batch of category-mistakes by showing that logically absurd corollaries follow from it. The exhibition of these absurdities will have the constructive effect of bringing out part of the correct logic of mental-conduct concepts.

4. Historical Note

It would not be true to say that the official theory derives solely from Descartes' theories, or even from a more widespread anxiety about the implications of seventeenth century mechanism. Scholastic and Reformation theology had schooled the intellects of the scientists as well as of the laymen, philosophers and clerics of that age. Stoic-Augustinian theories of the will were embedded in the Calvinist doctrines of sin and grace; Platonic and Aristotelian theories of the intellect shaped the orthodox doctrines of the immortality of the soul. Descartes was reformulating already prevalent theological doctrines of the soul in the new syntax of Galileo. The theologian's privacy of conscience became the philosopher's privacy of consciousness, and what had been the bogy of Predestination reappeared as the bogy of Determinism.

It would also not be true to say that the two-worlds myth did no theoretical good. Myths often do a lot of theoretical good, while they are still new. One benefit bestowed by the para-mechanical myth was that it partly superannuated the then prevalent para-political myth. Minds and their Faculties had previously been described by analogies with political superiors and political subordinates. The idioms used were those of ruling, obeying, collaborating and rebelling. They survived and still survive in many ethical and some epistemological discussions. As, in physics, the new myth of occult Forces was a scientific improvement on the old myth of Final Causes, so, in anthropological and psychological theory, the new myth of hidden operations, impulses and agencies was an improvement on the old myth of dictations, deferences and disobedience.

The Logical Analysis of Psychology

Carl G. Hempel

I

One of the most important and most discussed problems of contemporary philosophy is that of determining how psychology should be characterized in the theory of science. This problem, which reaches beyond the limits of epistemological analysis and has engendered heated controversy in metaphysics itself, is brought to a focus by the familiar alternative, 'Is psychology a natural science, or is it one of the sciences of mind and culture (*Geisteswissenschaften*)?'

The present article attempts to sketch the general lines of a new analysis of psychology, one which makes use of rigorous logical tools, and which has made possible decisive advances toward the solution of the above problem.[1] This analysis was carried out by the 'Vienna Circle', the members of which (M. Schlick, R. Carnap, P. Frank, O. Neurath, F. Waismann, H. Feigl, etc.) have, during the past ten years, developed an extremely fruitful method for the epistemological examination and critique of the various sciences, based in part on the work of L. Wittgenstein.[2] We shall limit ourselves essentially to the examination of psychology as carried out by Carnap and Neurath.

The method characteristic of the studies of the Vienna Circle can be briefly defined as a *logical analysis of the language of science.* This method became possible only with the development of a subtle logical apparatus which makes use, in particular, of all the formal procedures of modern symbolic logic.[3] However, in the following account, which does not pretend to give more than a broad orientation, we shall limit ourselves to setting out the general principles of this new method, without making use of strictly formal procedures.

II

Perhaps the best way to characterize the position of the Vienna Circle as it relates to psychology, is to say that it is the exact antithesis of the current epistemological thesis that there is a fundamental difference between experimental psychology, a natural science, and introspective psychology; and in general, between the natural sciences on the one hand, and the sciences of mind and culture on the other.[4] The common content of the widely different formulations used to express this contention, which we reject, can be set down as follows. Apart from certain aspects clearly related to physiology, psychology is radically different, both in subject matter and in method, from physics in the broad sense of the term. In particular, it is impossible to deal adequately with the subject matter of psychology by means of physical methods. The subject matter of physics includes such concepts as mass, wave length, temperature, field intensity, etc. In dealing with these, physics employs its distinctive method which makes a combined use of description and causal explanation. Psychology, on the other hand, has for its subject matter notions which are, in a broad sense, mental. They are *toto genere* different from the concepts of physics, and the appropriate method for dealing with them scientifically is that of empathetic insight, called 'introspection,' a method which is peculiar to psychology.

One of the principal differences between the two kinds of subject matter is generally believed to consist in the fact that the objects investigated by psychology—in contradistinction to those of physics—are specifically endowed with meaning. Indeed, several proponents of this idea state that the distinctive

From Carl G. Hempel. "The Logical Analysis of Psychology." From Ned Block, ed., *Readings in Philosophy of Psychology* (Harvard University Press,1980), I:14–23. First published as "Analyse logique de la psychologie", *Revue de Synthèse* 9-10: 27–42, 1935.

First translated in Herbert Feigl and Wilfrid Sellars, eds., *Readings in Philosophical Analysis* (Appleton-Century-Crofts, 1949), pp. 373–84.

method of psychology consists in 'understanding the sense of meaningful structures' (*sinnvoile Gebilde ver stehend zu erfassen*). Take, for example, the case of a man who speaks. Within the framework of physics, this process is considered to be completely explained once the movements which make up the utterance have been traced to their causes, that is to say, to certain physiological processes in the organism, and, in particular, in the central nervous system. But, it is said, this does not even broach the psychological problem. The latter begins with understanding the sense of what was said, and proceeds to integrate it into a wider context of meaning.

It is usually this latter idea which serves as a principle for the fundamental dichotomy that is introduced into the classification of the sciences. There is taken to be an *absolutely impassable gulf* between the *natural sciences* which have a subject matter devoid of meaning and the *sciences of mind and culture*, which have an intrinsically meaningful subject matter, the appropriate methodological instrument for the scientific study of which is 'comprehension of meaning.'

III

The position in the theory of science which we have just sketched has been attacked from several different points of view.[5] As far as psychology is concerned, one of the principal countertheses is that formulated by behaviorism, a theory born in America shortly before the war. (In Russia, Pavlov has developed similar ideas.) Its principal methodological postulate is that a scientific psychology should limit itself to the study of the bodily behavior with which man and the animals respond to changes in their physical environment, and should proscribe as nonscientific any descriptive or explanatory step which makes use of terms from introspective or 'understanding' psychology, such as 'feeling,' 'lived experience,' 'idea,' 'will,' 'intention,' 'goal,' 'disposition,' 'repress ion.'[6] We find in behaviorism, consequently, an attempt to construct a scientific psychology which would show by its success that even in psychology we have to do with purely physical processes, and that therefore there can be no impassable barrier between psychology and physics. However, this manner of undertaking the critique of a scientific thesis is not completely satisfactory. It seems, indeed, that the soundness of the behavioristic thesis expounded above depends on the possibility of fulfilling the program of behavioristic psychology. But one cannot expect the question as to the scientific status of psychology to be settled by empirical research in psychology itself. To achieve this is rather an undertaking in epistemology. We turn, therefore, to the considerations advanced by members of the Vienna Circle concerning this problem.

IV

Before addressing the question whether the subject matters of physics and psychology are essentially the same or different in nature, it is necessary first to clarify the very concept of the subject matter of a science. The theoretical content of a science is to be found in statements. It is necessary, therefore, to determine whether there is a fundamental difference between the statements of psychology and those of physics. Let us therefore ask what it is that determines the content—one can equally well say the 'meaning'—of a statement. When, for example, do we know the meaning of the following statement: 'Today at one o'clock, the temperature of such and such a place in the physics laboratory was 23.4° centigrade'? Clearly when, and only when, we know under what conditions we would call the statement true, and under what circumstances we would call it false. (Needless to say, it is not necessary to know whether or not the statement is true.) Thus, we understand the meaning of the above statement since we know that it is true when a tube of a certain kind filled with mercury (in short, a thermometer with a centigrade scale), placed at the indicated time at the location in question, exhibits a coincidence between the level of the mercury and the mark of the scale numbered 23.4. It is also true if in the same circumstances one can observe certain coincidences on another instrument called an 'alcohol thermometer'; and, again, if a galvanometer connected with a thermopile shows a certain deviation when the thermopile is placed there at the indicated time. Further, there is a long series of other possibilities which make the statement true, each of which is described by a 'physical test sentence,' as we will call it. The statement itself clearly affirms nothing other than this: all these physical test sentences obtain. (However, one verifies only some of these physical test sentences, and then 'concludes by induction' that the others obtain as well.) The

statement, therefore, is nothing but an abbreviated formulation of all those test sentences.

Before continuing the discussion, let us sum up this result as follows:

1. A statement that specifies the temperature at a selected point in space-time can be 'retranslated' without change of meaning into another statement—doubtless longer—in which the word 'temperature' no longer appears. That term functions solely as an abbreviation, making possible the concise and complete description of a state of affairs the expression of which would otherwise be very complicated.
2. The example equally shows that *two statements which differ in formulation* can nevertheless have the *same meaning*. A trivial example of a statement having the same meaning as the above would be: 'Today at one o'clock, at such and such a location in the laboratory, the temperature was 19.44° Réaumur.'

As a matter of fact, the preceding considerations show—and let us set it down as another result—that *the meaning of a statement is established by the conditions of its verification.* In particular, two differently formulated statements have the same meaning or the same effective content when, and only when, they are both true or both false in the same conditions. Furthermore, a statement for which one can indicate absolutely no conditions which would verify it, which is in principle incapable of confrontation with test conditions, is wholly devoid of content and without meaning. In such a case we have to do, not with a statement properly speaking, but with a 'pseudo-statement,' that is to say, a sequence of words correctly constructed from the point of view of grammar, but without content.[7]

In view of these considerations, our problem reduces to one concerning the difference between the circumstances which verify psychological statements and those which verify the statements of physics. Let us therefore examine a statement which involves a psychological concept, for example: 'Paul has a toothache.' What is the specific content of this statement, that is to say, what are the circumstances in which it would be verified? It will be sufficient to indicate some test sentences which describe these circumstances.

a. Paul weeps and makes gestures of such and such kinds.
b. At the question 'What is the matter?' Paul utters the words 'I have a toothache.'
c. Closer examination reveals a decayed tooth with exposed pulp.
d. Paul's blood pressure, digestive processes, the speed of his reactions, show such and such changes.
e. Such and such processes occur in Paul's central nervous system.

This list could be expanded considerably, but it is already sufficient to bring out the fundamental and essential point, namely, that all the circumstances which verify this psychological statement are expressed by physical test sentences. [This is true even of test condition *b*, which merely expresses the fact that in specified physical circumstances (the propagation of vibrations produced in the air by the enunciation of the words, 'What is the matter?') there occurs in the body of the subject a certain physical process (speech behavior of such and such a kind).]

The statement in question, which is about someone's 'pain,' is therefore, just like that concerning the temperature, simply an abbreviated expression of the fact that all its test sentences are verified.[8] (Here, too, one verifies only some of the test sentences and then infers by way of induction that the others obtain as well.) It can be retranslated without loss of content into a statement which no longer contains the term 'pain,' but only physical concepts. Our analysis has consequently established that a certain statement belonging to psychology has the same content as a statement belonging to physics; a result which is in direct contradiction to the thesis that there is an impassable gulf between the statements of psychology and those of physics.

The above reasoning can be applied to *any psychological statement*, even to those which concern, as is said, 'deeper psychological strata' than that of our example. Thus, the assertion that Mr. Jones suffers from intense inferiority feelings of such and such kinds can be confirmed or falsified only by observing Mr. Jones' behavior in various circumstances. To this behavior belong all the bodily processes of Mr. Jones, and, in particular, his gestures, the flushing and paling of his skin, his utterances, his blood pressure, the events that occur in his central nervous system, etc. In practice, when one wishes to test statements concerning what are called the deeper layers of the psyche, one limits oneself to the observation of external bodily behavior, and, particularly, to speech movements evoked by certain physical stimuli

(the asking of questions). But it is well known that experimental psychology has also developed techniques for making use of the subtler bodily states referred to above in order to confirm the psychological discoveries made by cruder methods. The statement concerning the inferiority feelings of Mr. Jones—whether true or false—means only this: such and such happenings take place in Mr. Jones' body in such and such circumstances.

We shall call a statement which can be translated without change of meaning into the language of physics, a 'physicalistic statement,' whereas we shall reserve the expression 'statement of physics' to those which are already formulated in the terminology of physical science. (Since every statement is in respect of content equivalent to itself, every statement of physics is also a physicalistic statement.) The result of the preceding considerations can now be summed up as follows: *All psychological statements which are meaningful, that is to say, which are in principle verifiable, are translatable into statements which do not involve psychological concepts, but only the concepts of physics. The statements of psychology are consequently physicalistic statements. Psychology is an integral part of physics.* If a distinction is drawn between psychology and the other areas of physics, it is only from the point of view of the practical aspects of research and the direction of interest, rather than a matter of principle. This logical analysis, the result of which shows a certain affinity with the fundamental ideas of behaviorism, constitutes the physicalistic conception of psychology.

V

It is customary to raise the following fundamental objection against the above conception. The physical test sentences of which you speak are absolutely incapable of formulating the intrinsic nature of a mental process; they merely describe the physical *symptoms* from which one infers, by purely psychological methods—notably that of understanding—the presence of a certain mental process.

But it is not difficult to see that the use of the method of understanding or of other psychological procedures is bound up with the existence of certain observable physical data concerning the subject undergoing examination. There is no psychological understanding that is not tied up physically in one way or another with the person to be understood. Let us add that, for example, in the case of the statement about the inferiority complex, even the 'introspective' psychologist, the psychologist who 'understands,' can confirm his conjecture only if the body of Mr. Jones, when placed in certain circumstances (most frequently, subjected to questioning), reacts in a specified manner (usually, by giving certain answers). Consequently, even if the statement in question had to be arrived at, *discovered*, by 'empathetic understanding,' the only *information* it gives us is nothing more nor less than the following: under certain circumstances, certain specific events take place in the body of Mr. Jones. It is this which constitutes the meaning of the psychological statement.

The further objection will perhaps be raised that men can feign. Thus, though a criminal at the bar may show physical symptoms of mental disorder, one would nevertheless be justified in wondering whether his mental confusion was 'real' or only simulated. One must note that in the case of the simulator, only some of the conditions are fulfilled which verify the statement 'This man is mentally unbalanced,' those, namely, which are most accessible to direct observation. A more penetrating examination—which should in principle take into account events occurring in the central nervous system—would give a decisive answer; and this answer would in turn clearly rest on a physicalistic basis. If, at this point, one wished to push the objection to the point of admitting that a man could show *all the 'symptoms'* of a mental disease without being 'really' ill, we reply that it would be absurd to characterize such a man as 'really normal'; for it is obvious that by the very nature of the hypothesis we should possess no criterion in terms of which to distinguish this man from another who, while exhibiting the same bodily behavior down to the last detail, would 'in addition' be 'really ill.' (To put the point more precisely, one can say that this hypothesis contains a *logical contradiction*, since it amounts to saying, 'It is possible that a statement should be false even when the necessary and sufficient conditions of its truth are fulfilled.')

Once again we see clearly that the meaning of a psychological statement consists solely in the function of abbreviating the description of certain modes of physical response characteristic of the bodies of men or animals. An analogy suggested by O. Neurath may be of further assistance in clarifying the logical function of psychological statements.[9] The complicated statements that would describe the movements of the hands of a watch in relation to one

another, and relatively to the stars, are ordinarily summed up in an assertion of the following form: 'This watch runs well (runs badly, etc.).' The term 'runs' is introduced here as an auxiliary defined expression which makes it possible to formulate briefly a relatively complicated system of statements. It would thus be absurd to say, for example, that the movement of the hands is only a 'physical symptom' which reveals the presence of a running which is intrinsically incapable of being grasped by physical means, or to ask, if the watch should stop, what has become of the running of the watch.

It is in exactly the same way that abbreviating symbols are introduced into the language of physics, the concept of temperature discussed above being an example. The system of physical test sentences *exhausts* the meaning of the statement concerning the temperature at a place, and one should not say that these sentences merely have to do with 'symptoms' of the existence of a certain temperature.

Our argument has shown that it is necessary to attribute to the characteristic concepts of psychology the same logical function as that performed by the concepts of 'running' and of 'temperature.' They do nothing more than make possible the succinct formulation of propositions concerning the states or processes of animal or human bodies.

The introduction of new psychological concepts can contribute greatly to the progress of scientific knowledge. But it is accompanied by a danger, that, namely, of making an excessive and, consequently, improper use of new concepts, which may result in questions and answers devoid of sense. This is frequently the case in metaphysics, notably with respect to the notions which we formulated in section II. Terms which are abbreviating symbols are imagined to designate a special class of 'psychological objects,' and thus one is led to ask questions about the 'essence' of these objects, and how they differ from 'physical objects.' The time-worn problem concerning the relation between mental and physical events is also based on this confusion concerning the logical function of psychological concepts. Our argument, therefore, enables us to see that *the psycho-physical problem is a pseudo-problem*, the formulation of which is based on an inadmissible use of scientific concepts; it is of the same logical nature as the question, suggested by the example above, concerning the relation of the running of the watch to the movement of the hands.[10]

VI

In order to bring out the exact status of the fundamental idea of the physicalistic conception of psychology (or logical behaviorism), we shall contrast it with certain theses of psychological behaviorism and of classical materialism, which give the appearance of being closely related to it.[11]

1. Logical behaviorism claims neither that minds, feelings, inferiority complexes, voluntary actions, etc., do not exist, nor that their existence is in the least doubtful. It insists that the very question as to whether these psychological constructs really exist is already a pseudo-problem, since these notions in their 'legitimate use' appear only as abbreviations in physicalistic statements. Above all, one should not interpret the position sketched in this paper as amounting to the view that we can know only the 'physical side' of psychological processes, and that the question whether there are mental phenomena behind the physical processes falls beyond the scope of science and must be left either to faith or to the conviction of each individual. On the contrary, the logical analyses originating in the Vienna Circle, one of whose consequences is the physicalistic conception of psychology, teach us that every meaningful question is, in principle, capable of a scientific answer. Furthermore, these analyses show that what, in the case of the mind-body problem, is considered as an object of belief, is absolutely incapable of being expressed by a factual proposition. In other words, there can be no question here of an 'article of faith.' Nothing can be an object of faith which cannot, in principle, be an object of knowledge.
2. The thesis here developed, though related in certain ways to the fundamental idea of behaviorism, does not demand, as does the latter, that psychological research restrict itself methodologically to the study of the responses organisms make to certain stimuli. It by no means offers a theory belonging to the domain of psychology, but rather a logical theory about the statements of scientific psychology. Its position is that the latter are without exception physicalistic statements, by whatever means they may have been

obtained. Consequently, it seeks to show that if in psychology only physicalistic statements are made, this is not a limitation because it is logically *impossible* to do otherwise.

3. In order for logical behaviorism to be valid, it is not necessary that we be able to describe the physical state of a human body which is referred to by a certain psychological statement—for example, one dealing with someone's feeling of pain—down to the most minute details of the phenomena of the central nervous system. No more does it presuppose a knowledge of all the physical laws governing human or animal bodily processes; nor *a fortiori* is the existence of rigorously deterministic laws relating to these processes a necessary condition of the truth of the behavioristic thesis. At no point does the above argument rest on such a concrete presupposition.

VII

In concluding, I should like to indicate briefly the clarification brought to the problem of the division of the sciences into totally different areas, by the method of the logical analysis of scientific statements, applied above to the special case of the place of psychology among the sciences. The considerations we have advanced can be extended to the domain of sociology, taken in the broad sense as the science of historical, cultural, and economic processes. In this way one arrives at the result that every sociological assertion which is meaningful, that is to say, in principle verifiable, 'has as its subject matter nothing else than the states, processes and behavior of groups or of individuals (human or animal), and their responses to one another and to their environment,'[12] and consequently that every sociological statement is a physicalistic statement. This view is characterized by Neurath as the thesis of 'social behaviorism,' which he adds to that of 'individual behaviorism' which we have expounded above. Furthermore, it can be shown[13] that every statement of what are called the 'sciences of mind and culture' is a sociological statement in the above sense, provided it has genuine content. Thus one arrives at the 'thesis of the unity of science':

The division of science into different areas rests exclusively on differences in research procedures and direction of interest; one must not regard it as a matter of principle. On the contrary, all the branches of science are in principle of one and the same nature; they are branches of the unitary science, physics.

VIII

The method of logical analysis which we have attempted to explicate by clarifying, as an example, the statements of psychology, leads, as we have been able to show only too briefly for the sciences of mind and culture, to a 'physicalism' based on logic (Neurath): *Every statement of the above-mentioned disciplines, and, in general, of empirical science as a whole*, which is not merely a meaningless sequence of words, *is translatable, without change of content, into a statement containing only physicalistic terms, and consequently is a physicalistic statement.*

This thesis frequently encounters strong opposition arising from the idea that such analyses violently and considerably reduce the richness of the life of mind or spirit, as though the aim of the discussion were purely and simply to eliminate vast and important areas of experience. Such a conception comes from a false interpretation of physicalism, the main elements of which we have already examined in section VII above. As a matter of fact, nothing can be more remote from a philosophy which has the methodological attitude we have characterized than the making of decisions, on its own authority, concerning the truth or falsity of particular scientific statements, or the desire to eliminate any matters of fact whatsoever. *The subject matter of this philosophy is limited to the form of scientific statements, and the deductive relationships obtaining between them.* It is led by its analyses to the thesis of physicalism, and establishes on purely logical grounds that a certain class of venerable philosophical 'problems' consists of pseudo-problems. It is certainly to the advantage of the progress of scientific knowledge that these imitation jewels in the coffer of scientific problems be known for what they are, and that the intellectual powers which have till now been devoted to a class of meaningless questions which are by their very nature insoluble, become available for the formulation and study of new and fruitful problems. That the method of logical analysis stimulates research along these lines is shown by the numerous publications of the Vienna Circle and those who

sympathize with its general point of view (H. Reichenbach, W. Dubislav, and others).

In the attitude of those who are so bitterly opposed to physicalism, an essential role is played by certain psychological factors relating to individuals and groups. Thus the contrast between the constructs (*Gebilde*) developed by the psychologist, and those developed by the physicist, or, again, the question as to the nature of the specific subject matter of psychology and the cultural sciences (which present the appearance of a search for the essence and unique laws of 'objective spirit') is usually accompanied by a strong emotional coloring which has come into being during the long historical development of a 'philosophical conception of the world,' which was considerably less scientific than normative and intuitive. These emotional factors are still deeply rooted in the picture by which our epoch represents the world to itself. They are protected by certain affective dispositions which surround them like a rampart, and for all these reasons appear to us to have genuine content—something which a more penetrating analysis shows to be impossible.

A psychological and sociological study of the causes for the appearance of these 'concomitant factors' of the metaphysical type would take us beyond the limits of this study,[14] but without tracing it back to its origins, it is possible to say that if the logical analyses sketched above are correct, the fact that they necessitate at least a partial break with traditional philosophical ideas which are deeply dyed with emotion can certainly not justify an opposition to physicalism—at least if one acknowledges that philosophy is to be something more than the expression of an individual vision of the world, that it aims at being a science.

NOTES

1. I now consider the type of physicalism outlined in this paper as too restrictive; the thesis that all statements of empirical science are *translatable*, without loss of theoretical content, into the language of physics, should be replaced by the weaker assertion that all statements of empirical science are *reducible* to sentences in the language of physics, in the sense that for every empirical hypothesis, including, of course, those of psychology, it is possible to formulate certain test conditions in terms of physical concepts which refer to more or less directly observable physical attributes. But those test conditions are not asserted to exhaust the theoretical content of the given hypothesis in all cases. For a more detailed development of this thesis, cf. R. Carnap, "Logical Foundations of the Unity of Science," reprinted in A. Marras, ed., *Intentionally. Mind, and Language* (Urbana: University of Illinois Press, 1972).
2. Ludwig Wittgenstein, *Tractatus Logico-Philosophicus* (London: Kegan Paul, 1922).
3. A recent presentation of symbolic logic, based on the fundamental work of Whitehead and Russell, *Principia Mathematica*, is to be found in R. Carnap, *Abriss der Logistik* (Vienna: Springer, 1929); vol. 2 of the series *Schriften zur Wissenschaftlichen Weltauffassung*. It includes an extensive bibliography, as well as references to other logistic systems.
4. The following are some of the principal publications of the Vienna Circle on the nature of psychology as a science: R. Carnap, *Scheinprobleme in der Philosophie. Das Fremdpsychische und des Realismusstreit* (Leipzig: Meiner, 1928); Rudolph Carnap, *Der Logische Aufbau der Welt* (Leipzig: Meiner, 1928), (English trans. *Logical Structure of the World* [Berkeley: University of California Press, 1967]); R. Carnap, Die Physikalische Sprache Als Universalsprache Der Wissenschaft *Erkenntnis* 2 (1931–32): pp. 432–65 [English trans.: "The Unity of Science" (London: Kegan Paul, 1934)]; "Psychologie in physikalischer Sprache," *Erkenntnis* 3 (1932–33), pp. 107–42; [English trans.: "Psychology in Physical Language," in A. J. Ayer, ed., *Logical Positivism* (New York: Free Press, 1959)]; "Ueber Protokollsaetze," *Erkenntnis* 3 (1932–33): pp. 215–28; O. Neurath, "Protokollsaetze," *Erkenntnis* 3 (1932–33): pp. 204–14 [English trans. "Protocol Sentences," in *Logical Positivism*]; *Einheitswissenschaft und Psychologie* (Vienna: Springer, 1933); vol. 1 of the series *Einheitswissenschaft*). See also the publications mentioned in the notes below.
5. P. Oppenheim, for example, in his book *Die Natuerliche Ordnung der Wissenschaften* (Jena: Fischer, 1926), opposes the view that there are fundamental differences between any of the different areas of science. On the analysis of 'understanding,' cf. M. Schlick, "Erleben, Erkennen, Metaphysik," *Kantstudien* 31 (1926): p. 146.
6. For further details, see the statement of one of the founders of behaviorism: J. B. Watson, *Behaviorism* (New York: Norton, 1930); also A. A. Roback, *Behaviorism and Psychology* (Cambridge, MA: University Bookstore, 1923); and A. P. Weiss, *A Theoretical Basis of Human Behavior*, 2nd ed., rev. (Columbus, OH: Adams, 1929); see also the work by Koehler cited in note 11 below.
7. Space is lacking for further discussion of the logical form of test sentences (recently called 'protocol sentences' by Neurath and Carnap). On this question, see Wittgenstein, *Tractatus Logico-Philosophicus*, as well as the articles by Neurath and Carnap, which have appeared in *Erkenntnis* (see note 4).
8. Two critical comments, 1977: (a) This reference to verification involves a conceptual confusion. The

thesis that the preceding considerations were intended to establish was clearly that the statement 'Paul has a toothache' is, in effect, an abbreviated expression of all its test sentences; not that it expresses the claim (let alone the 'fact') that all those test sentences have actually been tested and verified. (b) Strictly speaking, none of the test sentences just mentioned is implied by the statement 'Paul has a toothache': the latter may be true and yet any or all of those test sentences may be false. Hence, the preceding considerations fail to show that the given psychological statement can be 'translated' into sentences that, in purely physical terms, describe macro-behavioral manifestations of pain. This failure of the arguments outlined in the text does not preclude the possibility, however, that sentences ascribing pain or other psychological characteristics to an individual might be 'translatable,' in a suitable sense, into physical sentences ascribing associated physical micro-states or micro-events to the nervous system or to the entire body of the individual in question.

9. "Soziologie im Physikalismus," *Erkenntnis* 2 (1931–32): pp. 393–431, particularly p. 411 (English trans. "Sociology and Physicalism," in A.J. Ayer, ed., *Logical Positivism*).
10. R. Carnap, *Der Logische Aufbau der Welt*, pp. 231–36; *id.* Scheinprobleme in der Philosophie. See also note 4 above.
11. A careful discussion of the ideas of so-called 'internal' behaviorism is to be found in *Psychologische Probleme* by W. Koehler (Berlin: Springer, 1933). See particularly the first two chapters.
12. R. Carnap, "Die Physikalische Sprache als Universalsprache," p. 451. See also O. Neurath, *Empirische Soziologie* (Vienna: Springer, 1931); the fourth monograph in the series *Schriften zur wissenschaftlichen Weltauffassung*.
13. See R. Carnap, *Der Logische Aufbau der Welt*, pp. 22–34 and 185–211, as well as the works cited in the preceding note.
14. O. Neurath has made interesting contributions along these lines in *Empirische Soziologie* and in "Soziologie im Physikalismus" (see above, note 9), as has R. Carnap in his article "Ueberwindung der Metaphysik durch logische Analyse der Sprache," *Erkenntnis* 2 (1931–32): pp. 219–41 (English trans. "The Elimination of Metaphysics through Logical Analysis of Language," in A. J. Ayer, ed., *Logical Positivism*).

Brains and Behaviour[1]

Hilary Putnam

Once upon a time there was a tough-minded philosopher who said, 'What is all this talk about 'minds,' 'ideas,' and 'sensations'? Really—and I mean *really* in the real world—there is nothing to these so-called 'mental' events and entities but certain processes in our all-too-material heads.'

And once upon a time there was a philosopher who retorted, 'What a masterpiece of confusion! Even if, say, *pain* were perfectly correlated with any particular event in my brain (which I doubt) that event would obviously have certain properties—say, a certain numerical intensity measured in volts—which it would be *senseless* to ascribe to the feeling of pain. Thus, it is *two* things that are correlated, not *one*—and to call *two* things *one* thing is worse than being mistaken; it is utter contradiction.'

For a long time dualism and materialism appeared to exhaust the alternatives. Compromises were attempted ('double aspect' theories), but they never won many converts and practically no one found them intelligible. Then, in the mid-1930s, a seeming third possibility was discovered. This third possibility has been called *logical behaviourism.* To state the nature of this third possibility briefly, it is necessary to recall the treatment of the natural numbers (i.e., zero, one, two, three . . .) in modern logic. Numbers are identified with *sets*, in various ways, depending on which authority one follows. For instance, Whitehead and Russell identified zero with the set of all empty sets, one with the set of all one-membered sets, two with the set of all two-membered sets, three with the set of all three-membered sets, and so on. (This has the appearance of circularity, but they were able to dispel this appearance by defining 'one-membered set,' 'two-membered set,' 'three-membered set,' &c., without using 'one,' 'two,' 'three,' &c.) In short, numbers are treated as *logical constructions out of sets.* The number theorist is doing set theory without knowing it, according to this interpretation.

What was novel about this was the idea of getting rid of certain philosophically unwanted or embarrassing entities (numbers) without failing to do justice to the appropriate body of discourse (number theory) by treating the entities in question as logical constructions. Russell was quick to hold up this 'success' as a model to all future philosophers. And certain of those future philosophers—the Vienna positivists, in their 'physicalist' phase (about 1930)—took Russell's advice so seriously as to produce the doctrine that we are calling *logical behaviourism*—the doctrine that, just as numbers are (allegedly) logical constructions out of *sets*, so *mental events* are logical constructions out of actual and possible *behaviour events.*

In the set theoretic case, the 'reduction' of number theory to the appropriate part of set theory was carried out in detail and with indisputable technical success. One may dispute the philosophical significance of the reduction, but one knows exactly what one is talking about when one disputes it. In the mind-body case, the reduction was never carried out in even *one* possible way, so that it is not possible to be clear on just *how* mental entities or events are to be (identified with) logical constructions out of behaviour events. But, broadly speaking, it is clear what the view implies: it implies that all talk about mental events is translatable into talk about actual or potential overt behaviour.

It is easy to see in what way this view differs from both dualism and classical materialism. The logical behaviourist agrees with the dualist that what goes on in our brains has no connection whatsoever with what we *mean* when we say that someone is in pain. He can even take over the dualist's entire stock of arguments against the materialist position. Yet, at the same time, he can be as 'tough-minded' as the materialist in denying that ordinary talk of 'pains,' 'thoughts,' and 'feelings' involves reference to 'Mind' as a Cartesian substance.

Thus it is not surprising that logical behaviourism attracted enormous attention—both pro and con—during the next thirty years. Without doubt, this alternative proved to be a fruitful one to inject into the debate. Here, however, my

From R. Butler, ed., *Analytical Philosophy: Second Series* (Blackwell, 1968), pp. 1–19. Reprinted with permission of the publisher.

intention is not to talk about the fruitfulness of the investigations to which logical behaviourism has led, but to see if there was any upshot to those investigations. Can we, after thirty years, say anything about the rightness or wrongness of logical behaviourism? Or must we say that a third alternative has been added to the old two; that we cannot decide between three any more easily than we could decide between two; and that our discussion is thus half as difficult again as it was before?

One conclusion emerged very quickly from the discussion pro and con logical behaviourism: that the extreme thesis of logical behaviourism, as we just stated it (that all talk about 'mental events' is translatable into talk about overt behaviour) is false. But, in a sense, this is not very interesting. An extreme thesis may be false, although there is 'something to' the way of thinking that it represents. And the more interesting question is this: what, if anything, can be 'saved' of the way of thinking that logical behaviourism represents?

In the last thirty years, the original extreme thesis of logical behaviourism has gradually been weakened to something like this:

1. That there exist entailments between mind-statements and behaviour-statements; entailments that are not, perhaps, analytic in the way in which 'All bachelors are unmarried' is analytic, but that nevertheless follow (in some sense) from the meanings of mind words. I shall call these *analytic entailments.*
2. That these entailments may not provide an actual *translation* of 'mind talk' into 'behaviour talk' (this 'talk' talk was introduced by Gilbert Ryle in his *Concept of Mind*), but that this is true for such superficial reasons as the greater ambiguity of mind talk, as compared with the relatively greater specificity of overt behaviour talk.

I believe that, although no philosopher would today subscribe to the older version of logical behaviourism, a great many philosophers[2] would accept these two points, while admitting the unsatisfactory imprecision of the present statement of both of them. If these philosophers are right, then there is much work to be done (e.g., the notion of 'analyticity' has to be made clear), but the direction of work is laid out for us for some time to come.

I wish that I could share this happy point of view—if only for the comforting conclusion that first-rate philosophical research, continued for some time, will eventually lead to a solution to the mind-body problem which is independent of troublesome empirical facts about brains, central causation of behaviour, evidence for and against nonphysical causation of at least some behaviour, and the soundness or unsoundness of psychical research and parapsychology. But the fact is that I come to bury logical behaviourism, not to praise it. I feel that the time has come for us to admit that logical behaviourism is a mistake, and that even the weakened forms of the logical behaviourist doctrine are incorrect. I cannot hope to establish this in so short a paper as this one[3]; but I hope to expose for your inspection at least the main lines of my thinking.

Logical Behaviourism

The logical behaviourist usually begins by pointing out what is perfectly true, that such words as 'pain' ('pain' will henceforth be our stock example of a mind word) are not taught by reference to standard examples in the way in which such words as 'red' are. One can point to a standard red thing, but one cannot point to a standard pain (that is, except by pointing to some piece of *behaviour*) and say: 'Compare the feeling you are having with this one (say, Jones's feeling at time t_1). If the two feelings have the identical *quality*, then your feeling is legitimately called a feeling of *pain.*' The difficulty, of course, is that I cannot have Jones's feeling at time t_1—unless I *am* Jones, and the time *is* t_1.

From this simple observation, certain things follow. For example, the account according to which the *intension* of the word 'pain' is a certain *quality* which 'I know from my own case' must be wrong. But this is not to refute dualism, since the dualist need not maintain that I know the intension of the English word 'pain' from my own case, but only that I experience the referent of the word.

What then is the intension of 'pain'? I am inclined to say that 'pain' is a cluster-concept. That is, the application of the word 'pain' is controlled by a whole cluster of criteria, *all of which can be regarded as synthetic.*[4] As a consequence, there is no satisfactory way of answering the question 'What does 'pain' mean?' except by giving an exact synonym (e.g., 'Schmerz'); but there are a million and one different ways of saying what pain is. One can, for example, say that pain is that feeling which is

normally evinced by saying 'ouch,' or by wincing, or in a variety of other ways (or often not evinced at all).

All this is compatible with logical behaviourism. The logical behaviourist would reply: 'Exactly. 'Pain' is a cluster-concept—that is to say, it stands for *a cluster of phenomena.*' But that is not what I mean. Let us look at another kind of cluster-concept (cluster-concepts, of course, are not a homogeneous class): names of diseases.

We observe that, when a virus origin was discovered for polio, doctors said that certain cases in which all the symptoms of polio had been present, but in which the virus had been absent, had turned out not to be cases of polio at all. Similarly, if a virus should be discovered which normally (almost invariably) is the cause of what we presently call 'multiple sclerosis,' the hypothesis that this virus is *the* cause of multiple sclerosis would not be falsified if, in some few exceptional circumstances, it was possible to have all the symptoms of multiple sclerosis for some other combination of reasons, or if this virus caused symptoms not presently recognized as symptoms of multiple sclerosis in some cases. These facts would certainly lead the lexicographer to *reject* the view that 'multiple sclerosis' means 'the simultaneous presence of such and such symptoms.' Rather he would say that 'multiple sclerosis' means 'that disease which is normally responsible for some or all of the following symptoms. . . .'

Of course, he does not have to say this. Some philosophers would prefer to say that 'polio' *used to mean* 'the simultaneous presence of such-and-such symptoms.' And they would say that the *decision* to accept the presence or absence of a virus as a criterion for the presence or absence of polio represented a *change of meaning.* But this runs strongly counter to our common sense. For example, doctors used to say 'I believe polio is caused by a virus.' On the 'change of meaning' account, those doctors were *wrong*, not *right.* Polio, *as the word was then used*, was not always caused by a virus; it is only what *we* call polio that is always caused by a virus. And if a doctor ever said (many did) 'I believe this may not be a case of polio,' knowing that all of the text-book symptoms were present, that doctor must have been contradicting himself (even if we, to-day, would say that he was right) or, perhaps, 'making a disguised linguistic proposal.' Also, this account runs counter to good linguistic methodology. The definition we proposed a paragraph back—'multiple sclerosis' means 'the disease that is normally *responsible* for the following symptoms. . . .'—has an exact analogue in the case of polio. This kind of definition leaves open the question whether there is a single cause or several. It is consonant with such a definition to speak of 'discovering a single origin for polio (or two or three or four),' to speak of 'discovering X did not have polio' (although he exhibited all the symptoms of polio), and to speak of 'discovering X did have polio' (although he exhibited *none* of the 'textbook symptoms'). And, finally, such a definition does not require us to say that any 'change of meaning' took place. Thus, this is surely the definition that a good lexicographer would adopt. But this entails *rejecting* the 'change of meaning' account as a philosopher's invention.[5]

Accepting that this is the correct account of the names of diseases, what follows? There *may* be analytic entailments connecting diseases and symptoms (although I shall argue against this). For example, it looks plausible to say that:

> 'Normally people who have multiple sclerosis have some or all of the following symptoms . . .'

is a necessary ('analytic') truth. But it does not follow that 'disease talk' is translatable into 'symptom talk.' Rather the contrary follows (as is already indicated by the presence of the word 'normally'): statements about multiple sclerosis are not translatable into statements about the symptoms of multiple sclerosis, not because disease talk is 'systematically ambiguous' and symptom talk is 'specific,' but because *causes* are not logical constructions out of their *effects.*

In analogy with the foregoing, both the dualist and the materialist would want to argue that, although the meaning of 'pain' may be *explained* by reference to overt behaviour, what we mean by 'pain' is not the presence of a cluster of responses, but rather the presence of an event or condition that normally causes those responses. (Of course the pain is not the whole cause of the pain behaviour, but only a suitably invariant part of that cause[6] but, similarly, the virus-caused tissue damage is not the whole cause of the individual symptoms of polio in some individual case, but a suitably invariant part of the cause.) And they would want to argue further, that even if it *were* a necessary truth that

> 'Normally, when one says 'ouch' one has a pain'

or a necessary truth that

> 'Normally, when one has a pain one says 'ouch''

this would be an interesting observation about what 'pain' means, but it would shed no metaphysical light on what pain *is* (or *isn't*). And it certainly would not follow that 'pain talk' is translatable into 'response talk,' or that the failure of translatability is only a matter of the 'systematic ambiguity' of pain talk as opposed to the 'specificity' of response talk: quite the contrary. Just as before, *causes* (pains) are *not* logical constructions out of their *effects* (behaviour).

The traditional dualist would, however, want to go farther, and deny the *necessity* of the two propositions just listed. Moreover, the traditional dualist is right: there is nothing self-contradictory, as we shall see below, in talking of hypothetical worlds in which there are pains but *no* pain behaviour.

The analogy with names of diseases is still preserved at this point. Suppose I identify multiple sclerosis as the disease that normally produces certain symptoms. If it later turns out that a certain virus is the cause of multiple sclerosis, using this newly discovered criterion I may then go on to find out that multiple sclerosis has quite different symptoms when, say, the average temperature is lower. I can then perfectly well talk of a hypothetical world (with lower temperature levels) in which multiple sclerosis does *not* normally produce the usual symptoms. It is true that if the *words* 'multiple sclerosis' are used in any world in such a way that the above lexical definition is a good one, *then* many victims of the disease must have had some or all of the following symptoms . . . And in the same way it is true that *if* the explanation suggested of the word 'pain' is a good one (i.e., 'pain is the feeling that is normally being evinced when someone says 'ouch,' or winces, or screams, &c.'), *then* persons in pain must have at some time winced or screamed or said 'ouch'—but this does *not* imply that 'if someone ever had a pain, then someone must at some time have winced or screamed or said 'ouch.' To conclude this would be to confuse preconditions for *talking* about pain as *we* talk about pain with preconditions for the existence of pain.

The analogy we have been developing is not an identity: linguistically speaking, mind words and names of diseases are different in a great many respects. In particular, *first person uses* are very different: a man may have a severe case of polio and not know it, even if he knows the word 'polio,' but one cannot have a severe pain and not know it. At first blush, this may look like a point in favour of logical behaviourism. The logical behaviourist may say: it is because the premisses 'John says he has a pain,' 'John knows English,' and 'John is speaking in all sincerity,'[7] *entail* 'John has a pain,' that pain reports have this sort of special status. But even if this is right, it does not follow that logical behaviourism is correct unless *sincerity* is a 'logical construction out of overt behaviour'! A far more reasonable account is this: one can have a 'pink elephant hallucination,' but one cannot have a 'pain hallucination,' or an 'absence of pain hallucination,' simply because any situation that a person cannot discriminate from a situation in which he himself has a pain *counts* as a situation in which he has a pain, whereas a situation that a person cannot distinguish from one in which a pink elephant is present does not necessarily *count* as the presence of a pink elephant.

To sum up: I believe that pains are not clusters of responses, but that they are (normally, in our experience to date) the causes of certain clusters of responses. Moreover, although this is an empirical fact, it underlies the possibility of talking about pains in the particular way in which we do. However, it does not rule out in any way the possibility of worlds in which (owing to a difference in the environmental and hereditary conditions) pains are not responsible for the usual responses, or even are not responsible for any responses at all.

Let us now engage in a little science fiction. Let us try to describe some worlds in which pains are related to responses (and also to causes) in quite a different way than they are in our world.

If we confine our attention to non-verbal responses by full grown persons, for a start, then matters are easy. Imagine a community of 'super-spartans' or 'super-stoics'—a community in which the adults have the ability to successfully suppress *all* involuntary pain behaviour. They may, on occasion, admit that they feel pain, but always in pleasant well-modulated voices—even if they are undergoing the agonies of the damned. They do *not* wince, scream, flinch, sob, grit their teeth, clench their fists, exhibit beads of sweat, or otherwise act like people in pain or people suppressing the unconditioned responses associated with pain. However, they do feel pain, and they dislike it (just as we do). They even admit that it takes a great effort of will to behave as they do. It is only that they have what they regard as important ideological reasons for behaving as they do, and they have, through years of training, learned to live up to their own exacting standards.

It may be contended that children and not fully mature members of this community will exhibit, to varying degrees, normal unconditioned pain behaviour, and that this is all that is necessary for the ascription of pain. On this view, the sine qua non for the significant ascription of pain to a species is that its immature members should exhibit unconditioned pain responses.

One might well stop to ask whether this statement has even a clear meaning. Supposing that there are Martians: do we have any criterion for something being an 'unconditioned pain response' for a Martian? Other things being equal, one *avoids* things with which one has had painful experiences: this would suggest that *avoidance*-behaviour might be looked for as a universal unconditioned pain response. However, even if this were true, it would hardly be specific enough, since avoidance can also be an unconditioned response to many things that we do not associate with pain—to things that disgust us, or frighten us, or even merely bore us.

Let us put these difficulties aside, and see if we can devise an imaginary world in which there are not, even by lenient standards, any unconditioned pain responses. Specifically, let us take our 'super-spartans,' and let us suppose that after millions of years they begin to have children who are born fully acculturated. They are born speaking the adult language, knowing the multiplication table, having opinions on political issues, and inter alia sharing the dominant spartan beliefs about the importance of not evincing pain (except by way of a verbal report, and even that in a tone of voice that suggests indifference). Then there would not *be* any 'unconditioned pain responses' in this community (although there might be unconditioned *desires* to make certain responses—desires which were, however, always suppressed by an effort of will). Yet there is a clear absurdity to the position that one cannot ascribe to these people a capacity for feeling pain.

To make this absurdity evident, let us imagine that we succeed in converting an adult 'super-spartan' to *our* ideology. Let us suppose that he begins to evince pain in the normal way. Yet he reports that the pains he is feeling are not more *intense* than are the ones he experienced prior to conversion—indeed, he may say that giving expression to them makes them *less* intense. In this case, the logical behaviourist would have to say that, through the medium of this one member, we had demonstrated the existence of unconditioned pain responses in the whole species, and hence that ascription of pain to the species is 'logically proper.' But this is to say that had this one man never lived, and had it been possible to demonstrate only indirectly (via the use of *theories*) that these beings feel pain, then pain ascriptions *would* have been improper.

We have so far been constructing worlds in which the relation of pain to its non-verbal *effects* is altered. What about the relation of pain to *causes?* This is even more easy for the imagination to modify. Can one not imagine a species who feel pain only when a magnetic field is present (although the magnetic field causes no detectable damage to their bodies or nervous systems)? If we now let the members of such a species become converts *to* 'superspartanism,' we can depict to ourselves a world in which pains, in our sense, are clearly present, but in which they have neither the normal causes nor the normal effects (apart from verbal reports).

What about verbal reports? Some behaviourists have taken these as the characteristic form of pain behaviour. Of course, there is a difficulty here: If 'I am in pain' means 'I am disposed to utter this kind of verbal report' (to put matters crudely), then how do we tell that any particular report is 'this kind of verbal report'? The usual answer is in terms of the unconditioned pain responses and their assumed supplantation by the verbal reports in question. However, we have seen that there are no *logical* reasons for the existence of unconditioned pain responses in all species capable of feeling pain (there *may* be logical reasons for the existence of avoidance desires, but avoidance *desires* are not themselves behaviour any more than pains are).

Once again, let us be charitable to the extent of waving the first difficulty that comes to mind, and let us undertake the task of trying to imagine a world in which there are not even pain *reports.* I will call this world the 'X-world.' In the X-world we have to deal with 'super-super-spartans.' These have been super-spartans for so long, that they have begun to suppress even *talk* of pain. Of course, each individual X-worlder may have his private way of thinking about pain. He may even have the *word* 'pain' (as before, I assume that these beings are born fully acculturated). He may *think* to himself: 'This pain is intolerable. If it goes on one minute longer I shall scream. Oh No! I mustn't do that! That would disgrace my whole family . . .' But X-worlders do not even admit to *having* pains. They pretend not to know either

the word or the phenomenon to which it refers. In short, if pains are 'logical constructs out of behaviour,' then our X-worlders behave so as not to have pains!—Only, of course, they do have pains, and they know perfectly well that they have pains.

If this last fantasy is not, in some disguised way, self-contradictory, then logical behaviourism is simply a mistake. Not only is the second thesis of logical behaviourism—the existence of a near-translation of pain talk into behaviour talk—false, but so is even the first thesis—the existence of 'analytic entailments.' Pains *are* responsible for certain kinds of behaviour—but only in the context of our beliefs, desires, ideological attitudes, and so forth. From the statement 'X has a pain' by itself *no* behavioural statement follows—not even a behavioural statement with a 'normally' or a 'probably' in it.

In our concluding section we shall consider the logical behaviourist's stock of countermoves to this sort of argument. If the logical behaviourist's positive views are inadequate owing to an oversimplified view of the nature of cluster words—amounting, in some instances, to an open denial that it is *possible* to have a word governed by a cluster of indicators, *all* of which are synthetic—his negative views are inadequate owing to an oversimplified view of empirical reasoning. It is unfortunately characteristic of modern philosophy that its problems should overlap three different areas—to speak roughly, the areas of linguistics, logic, and 'theory of theories' (scientific methodology)—and that many of its practitioners should try to get by with an inadequate knowledge of at least two out of the three.

Some Behaviourist Arguments

We have been talking of 'X-worlders' and 'super-spartans.' No one denies that, in *some* sense of the term, such fantasies are 'intelligible.' But 'intelligibility' can be a superficial thing. A fantasy may be 'intelligible,' at least at the level of 'surface grammar,' although we may come to see, on thinking about it for a while, that some absurdity is involved. Consider, for example, the supposition that last night, just on the stroke of midnight, all distances were instantaneously doubled. Of course, we did not notice the change, for *we* ourselves also doubled in size! This story may seem intelligible to us at first blush, at least as an amusing possibility. On reflection, however, we come to see that a logical contradiction is involved. For 'length' means nothing more nor less than a relation to a standard, and it is a contradiction to maintain that the length of everything doubled, while the relations to the standards remained unchanged.

What I have just said (speaking as a logical behaviourist might speak) is false, but not totally so. It is false (or at least the last part is false), because 'length' does *not* mean 'relation to a standard.' If it did (assuming a 'standard' has to be a macroscopic material object, or anyway a material object), it would make no sense to speak of distances in a world in which there were only gravitational and electromagnetic fields, but no material objects. Also, it would make no sense to speak of the *standard* (whatever it might be) as having changed its length. Consequences so counter-intuitive have led many physicists (and even a few philosophers of physics) to view 'length' not as something operationally defined, but as a theoretical magnitude (like electrical charge), which can be measured in a virtual infinity of ways, but which is not explicitly and exactly definable in terms of any of the ways of measuring it. Some of these physicists—the 'unified field' theorists—would even say that, far from it being the case that 'length' (and hence 'space') depends on the existence of suitably related material bodies, material bodies are best viewed as local variations in the curvature of space—that is to say, local variations in the intensity of a certain magnitude (the tensor g_{ik}), one aspect of which we experience as 'length.'

Again, it is far from true that the hypothesis 'last night, on the stroke of midnight, everything doubled in length' has no testable consequences. For example, if last night everything did double in length, and the velocity of light did not also double, then this morning we would have experienced an apparent halving of the speed of light. Moreover, if g (the gravitational constant) did not double, then we would have experienced and apparent halving in the intensity of the gravitational field. And if h (Planck's constant) did not change, then. . . . In short, our world would have been bewilderingly different. And if we could survive at all, under so drastically altered conditions, no doubt some clever physicist would figure out what had happened.

I have gone into such detail just to make the point that in philosophy things are rarely so simple as they seem. The 'doubling universe' is a favourite classroom example of a 'pseudohypothesis'—yet it is the worst possible example if a 'clear case' is desired. In the first place,

what is desired is a hypothesis with no testable consequences—yet *this* hypothesis, as it is always stated, *does* have testable consequences (perhaps some more complex hypothesis does not; but then we have to see this more complex hypothesis stated before we can be expected to discuss it). In the second place, the usual argument for the absurdity of this hypothesis rests on a simplistic theory of the meaning of 'length'—and a full discussion of *that* situation is hardly possible without bringing in considerations from unified field theory and quantum mechanics (the latter comes in in connection with the notion of a 'material standard'). But, the example aside, one can hardly challenge the point that a superficially coherent story may contain a hidden absurdity.

Or can one? Of course, a superficially coherent story may contain a hidden logical contradiction, but the whole point of the logical behaviourist's sneering reference to 'surface grammar' is that *linguistic coherence, meaningfulness of the individual terms*, and *logical consistency*, do not by themselves guarantee freedom from another kind of absurdity—there are 'depth absurdities' which can only be detected by more powerful techniques. It is fair to say that to-day, after thirty years of this sort of talk, we lack both a single *convincing* example of such a depth absurdity, and a technique of detection (or alleged technique of detection) which does not reduce to 'untestable, *therefore* nonsense.'

To come to the case at hand: the logical behaviourist is likely to say that our hypothesis about 'X-worlders' is untestable in principle (if there *were* 'X-worlders,' by hypothesis we couldn't distinguish them from people who really didn't know what pain is); and *therefore* meaningless (apart from a certain 'surface significance' which is of no real interest). If the logical behaviourist has learned a little from 'ordinary language philosophy,' he is likely to shy away from saying 'untestable, therefore *meaningless*,' but he is still likely to say or at least think: 'untestable, therefore in *some* sense absurd.' I shall try to meet this 'argument' *not* by challenging the premiss, be it overt or covert, that 'untestable synthetic statement' is some kind of contradiction in terms (although I believe that premiss to be mistaken), but simply by showing that, on any but the most naive view of testability, our hypothesis *is* testable.

Of course, I could not do this if it were true that 'by hypothesis, we couldn't distinguish X-worlders from people who *really* didn't know what pain is.' But that isn't true—at any rate, it isn't true 'by hypothesis.' What is true by hypothesis is that we couldn't distinguish X-worlders from people who really didn't know what pain is *on the basis of overt behaviour alone*. But that still leaves many other ways in which we might determine what is going on 'inside' the X-worlders—in both the figurative and literal sense of 'inside.' For example, we might examine their *brains*.

It is a fact that when pain impulses are 'received' in the brain, suitable electrical detecting instruments record a characteristic 'spike' pattern. Let us express this briefly (and too simply) by saying that 'brain spikes' are one-to-one correlated with experiences of pain. If our X-worlders belong to the human species, then we can verify that they do feel pains, notwithstanding their claim that they don't have any idea what pain is, by applying our electrical instruments and detecting the tell-tale 'brain spikes.'

This reply to the logical behaviourist is far too simple to be convincing. 'It is true,' the logical behaviourist will object, 'that experiences of pain are one-to-one correlated with 'brain spikes' in the case of normal human beings. But you don't know that the X-worlders are normal human beings, in this sense—in fact, you have every reason to suppose that they are *not* normal human beings.' This reply shows that no *mere* correlation, however carefully verified in the case of normal human beings, can be used to verify ascriptions of pain to X-worlders. Fortunately, we do not have to suppose that our knowledge will always be restricted to mere correlations, like the pain–'brain spike' correlation. At a more advanced level, considerations of simplicity and coherence can begin to play a rôle in a way in which they cannot when only crude observational regularities are available.

Let us suppose that we begin to detect waves of a new kind, emanating from human brains—call them 'V-waves.' Let us suppose we develop a way of 'decoding' V-waves so as to reveal people's unspoken thoughts. And, finally, let us suppose that our 'decoding' technique also works in the case of the V-waves emanating from the brains of X-worlders. How does this correlation differ from the pain–'brain spike' correlation?

Simply in this way: it is reasonable to say that 'spikes'—momentary peaks in the electrical intensity in certain parts of the brain—could have almost any cause. But waves which go over into coherent English (or any other language); under

a relatively simple decoding scheme, could not have just any cause. The 'null hypothesis'—that this is just the operation of 'chance'—can be dismissed at once. And if, in the case of human beings, we verify that the decoded waves correspond to what we are in fact thinking, then the hypothesis that this same correlation holds in the case of X-worlders will be assigned an immensely high probability, simply because no other likely explanation readily suggests itself. But 'no other likely explanation readily suggests itself' isn't verification, the logical behaviourist may say. On the contrary. How, for example, have we verified that cadmium lines in the spectrographic analysis of sunlight indicate the presence of cadmium in the sun? Mimicking the logical behaviourist, we might say: 'We have verified that under normal circumstances, cadmium lines only occur when heated cadmium is present. But we don't know that circumstances on the sun are normal in this sense.' If we took this seriously, we would have to *heat cadmium on the sun* before we could say that the regularity upon which we base our spectrographic analysis of sunlight had been verified. In fact, we have verified the regularity under 'normal' circumstances, and we can *show* (deductively) that *if* many other laws, that have also been verified under 'normal' circumstances and *only* under 'normal' circumstances (i.e., never on the surface of the sun), hold on the sun, *then* this regularity holds also under 'abnormal' circumstances. And if someone says, 'But perhaps *none* of the usual laws of physics hold on the sun,' we reply that this is like supposing that a random process always produces coherent English. The fact is that the 'signals' (sunlight, radio waves, &c.) which we receive from the sun cohere with a vast body of theory. Perhaps there is some other explanation than that the sun obeys the usual laws of physics; but *no other likely explanation suggests itself.* This sort of reasoning *is* scientific verification; and if it is not reducible to simple Baconian induction—well, then, philosophers must learn to widen their notions of verification to embrace it.

The logical behaviourist might try to account for the decodability of the X-worlders' 'V-waves' into coherent English (or the appropriate natural language) without invoking the absurd 'null hypothesis.' He might suggest, for example, that the 'X-worlders' are having fun at our expense—they are able, say, to produce misleading V-waves at will. If the X-worlders have brains quite unlike ours, this may even have some plausibility. But once again, in an advanced state of knowledge, considerations of coherence and simplicity may quite conceivably 'verify' that this is false. For example, the X-worlders may have brains quite like ours, rather than unlike ours. And we may have built up enough theory to say how the brain of a human being should 'look' if that human being were pretending not to be in pain when he was, in fact, in pain. Now consider what the 'misleading V-waves' story requires: it requires that the X-worlders produce V-waves in quite a different way than we do, without specifying what that different way is. Moreover, it requires that this be the case, although the reverse hypothesis—that X-worlders' brains function *exactly* as human brains do—in fact, that they *are* human brains—fits all the data. Clearly, this story is in serious methodological difficulties, and any other 'counter-explanation' that the logical behaviourist tries to invoke will be in similar difficulties. In short, the logical behaviourist's argument reduces to this: 'You cannot verify 'psycho-physical' correlations in the case of X-worlders (or at least, you can't verify ones having to do, directly or indirectly, with *pain*), because, by hypothesis, X-worlders won't tell you (or indicate behaviourally) when they are in pain. 'Indirect verification'—verification using theories which have been 'tested' only in the case of human beings—is not verification at all, because X-worlders *may* obey different laws than human beings. And it is not incumbent upon *me* (the logical behaviourist says) to suggest what those laws might be: it is incumbent upon *you* to rule out *all* other explanations.' And this is a silly argument. The scientist does not have to rule out all the ridiculous theories that someone *might* suggest; he only has to show that he has ruled out any reasonable alternative theories that one might put forward on the basis of present knowledge.

Granting, then, that we might discover a technique for 'reading' the unspoken thoughts of X-worlders: we would then be in the same position with respect to the X-worlders as we were with respect to the original 'super-spartans.' The super-spartans were quite willing to tell us (and each other) about their pains; and we could see that their pain talk was linguistically coherent and situationally appropriate (e.g., a super-spartan will tell you that he feels intense pain when you touch him with a red hot poker). On this basis, we were quite willing to grant that the super-spartans did, indeed, feel pain—all the more readily, since the deviancy

in their behaviour had a perfectly convincing ideological explanation. (Note again the rôle played here by considerations of coherence and simplicity). But the X-worlders also 'tell' us (and, perhaps, each other), exactly the same things, albeit *un*willingly (by the medium of the involuntarily produced 'V-waves'). Thus we have to say—at least, we have to say as long as the 'V-wave' theory has not broken down—that the X-worlders are what they, in fact, are—just 'super-super-spartans.'

Let us now consider a quite different argument that a logical behaviourist might use. 'You are assuming,' he might say, 'the following principle:

> If someone's brain is in the same state as that of a human being in pain (not just at the moment of the pain, but before and after for a sufficient interval), then he is in pain.'

'Moreover, this principle is one which it would never be reasonable to give up (on your conception of 'methodology'). Thus, you have turned it into a tautology. But observe what turning this principle into a tautology involves: it involves changing the meaning of 'pain.' What 'pain' means for *you* is: the presence of pain, in the colloquial sense of the term, *or* the presence of a brain state identical with the brain state of someone who feels pain. Of course, in that sense we can verify that your 'X-worlders' experience 'pain'—but that is not the sense of 'pain' at issue.'

The reply to this argument is that the premiss is simply false. It is just not true that, on my conception of verification, it would *never* be reasonable to give up the principle stated. To show this, I have to beg your pardons for engaging in a little more science fiction. Let us suppose that scientists discover yet another kind of waves—call them 'W-waves.' Let us suppose that W-waves do not emanate from human brains, but that they are detected emanating from the brains of X-worlders. And let us suppose that, once again, there exists a simple scheme for decoding W-waves into coherent English (or whatever language X-worlders speak), and that the 'decoded' waves 'read' like this: 'Ho, ho! Are we fooling those Earthians! They think that the V-waves they detect represent our thoughts! If they only knew that instead of pretending not to have pains when we really have pains, we are really pretending to pretend not to have pains when we really do have pains when we really don't have pains!' Under these circumstances, we would 'doubt' (to put it mildly) that the same psycho-physical correlations held for normal humans and for X-worlders. Further investigations might lead us to quite a number of different hypotheses. For example, we might decide that X-worlders don't think with their brains at all—that the 'organ' of thought is not just the brain, in the case of X-worlders, but some larger structure—perhaps even a structure which is not 'physical' in the sense of consisting of elementary particles. The point is that what is necessarily true is not the principle stated two paragraphs back, but rather the principle:

> If someone (some organism) is in the same state as a human being in pain in all relevant respects, then he (that organism) is in pain.

—And *this* principle *is* a tautology by anybody's lights! The only a priori methodological restriction I am imposing here is this one:

> If some organism is in the same state as a human being in pain in all respects *known* to be relevant, and there is no reason to suppose that there exist *un*known relevant respects, then don't postulate any.

—But this principle is not a 'tautology'; in fact, it is not a *statement* at all, but a methodological directive. And deciding to conform to this directive is not (as hardly needs to be said) changing the meaning of the word 'pain,' or of *any* word.

There are two things that the logical behaviourist can do: he can claim that ascribing pains to X-worlders, or even super-spartans, involves a 'change of meaning,'[8] or he can claim that ascribing pains to super-spartans, or at least to X-worlders, is 'untestable.' The first thing is a piece of unreasonable linguistics; the second, a piece of unreasonable scientific method. The two are, not surprisingly, mutually supporting: the unreasonable scientific method makes the unreasonable linguistics appear more reasonable. Similarly, the normal ways of thinking and talking are mutually supporting: reasonable linguistic field techniques are, needless to say, in agreement with reasonable conceptions of scientific method. Madmen sometimes have consistent delusional systems; so madness and sanity can both have a 'circular' aspect. I may not have succeeded, in this paper, in breaking the 'delusional system' of a committed logical behaviourist; but I hope to have convinced the uncommitted that that system need not be taken seriously. If we have to choose between 'circles,' the circle of reason is to be preferred to any of the many circles of unreason.

NOTES

1. This paper was read as a part of the programme of the American Association for the Advancement of Science, Section L (History and Philosophy of Science), December 27, 1961.
2. E.g., these two points are fairly explicitly stated in Strawson's *Individuals.* Strawson has told me that he no longer subscribes to point (1), however.
3. An attempted fourth alternative—i.e., an alternative to dualism, materialism, *and* behaviourism—is sketched in "The Mental Life of Some Machines," which appeared in the Proceedings of the Wayne Symposium on the Philosophy of Mind. This fourth alternative is materialistic in the wide sense of being compatible with the view that organisms, including human beings, are physical systems consisting of elementary particles and obeying the laws of physics, but does not require that such 'states' as *pain* and *preference* be defined in a way which makes reference to either overt behaviour or physical-chemical constitution. The idea, briefly, is that predicates which apply to a system by virtue of its *functional organization* have just this characteristic: a given functional organization (e.g., a given inductive logic, a given rational preference function) may realize itself in almost any kind of overt behaviour, depending upon the circumstances, and is capable of being 'built into' structures of many different logically possible physical (or even metaphysical) constitutions. Thus the statement that a creature prefers A to B does not tell us whether the creature has a carbon chemistry, or a silicon chemistry, or is even a disembodied mind, nor does it tell us how the creature would behave under any circumstances specifiable without reference to the creature's other preferences and beliefs, but it does not thereby become something 'mysterious.'
4. I mean not only that *each* criterion can be regarded as synthetic, but also that the cluster is *collectively* synthetic, in the sense that we are free in certain cases to say (for reason of inductive simplicity and theoretical economy) that the term applies although the whole cluster is missing. This is completely compatible with saying that the cluster serves to fix the meaning of the word. The point is that when we specify something by a cluster of indicators we assume that people will *use their brains.* That criteria may be over-ridden when good sense demands is the sort of thing we may regard as a 'convention associated with discourse' (Grice) rather than as something to be stipulated in connection with the individual words.
5. Cf. "Dreaming and 'Depth Grammar,'" *Analytical Philosophy*, First Series.
6. Of course, 'the cause' is a highly ambiguous phrase. Even if it is correct in certain contexts to say that certain events in the brain are 'the cause' of my pain behaviour, it does *not* follow (as has sometimes been suggested) that my pain must be 'identical' with these neural events.
7. This is suggested in Wittgenstein's *Philosophical Investigations.*
8. This popular philosophical move is discussed in "Dreaming and 'Depth Grammar,'" *Analytical Philosophy*, First Series.

C. The Identity Theory and Functionalism

Sensations and Brain Processes

J. J. C. Smart

Suppose that I report that I have at this moment a roundish, blurry-edged after-image which is yellowish towards its edge and is orange towards its centre. What is it that I am reporting?[1] One answer to this question might be that I am not reporting anything, that when I say that it looks to me as though there is a roundish yellowy orange patch of light on the wall I am expressing some sort of *temptation*, the temptation to say that there *is* a roundish yellowy orange patch on the wall (though I may know that there is not such a patch on the wall). This is perhaps Wittgenstein's view in the *Philosophical Investigations* (see paragraphs 367, 370). Similarly, when I 'report' a pain, I am not really reporting anything (or, if you like, I am reporting in a queer sense of 'reporting'), but am doing a sophisticated sort of wince. (See paragraph 244: 'The verbal expression of pain replaces crying and does not describe it.' Nor does it describe anything else?)[2] I prefer most of the time to discuss an after-image rather than a pain, because the word 'pain' brings in something which is irrelevant to my purpose: the notion of 'distress.' I think that 'he is in pain' entails 'he is in distress,' that is, that he is in a certain agitation-condition.[3] Similarly, to say 'I am in pain' may be to do more than 'replace pain behavior': it may be partly to report something, though this something is quite nonmysterious, being an agitation-condition, and so susceptible of behavioristic analysis. The suggestion I wish if possible to avoid is a different one, namely that 'I am in pain' is a genuine report, and that what it reports is an irreducibly psychical something. And similarly the suggestion I wish to resist is also that to say 'I have a yellowish orange after-image' is to report something irreducibly psychical.

Why do I wish to resist this suggestion? Mainly because of Occam's razor. It seems to me that science is increasingly giving us a viewpoint whereby organisms are able to be seen as physico-chemical mechanisms:[4] it seems that even the behavior of man himself will one day be explicable in mechanistic terms. There does seem to be, so far as science is concerned, nothing in the world but increasingly complex arrangements of physical constituents. All except for one place: in consciousness. That is, for a full description of what is going on in a man you would have to mention not only the physical processes in his tissue, glands, nervous system, and so forth, but also his states of consciousness: his visual, auditory, and tactual sensations, his aches and pains. That these should be *correlated* with brain processes does not help, for to say that they are *correlated* is to say that they are something 'over and above.' You cannot correlate something with itself. You correlate footprints with burglars, but not Bill Sikes the burglar with Bill Sikes the burglar. So sensations, states of consciousness, do seem to be the one sort of thing left outside the physicalist picture, and for various reasons I just cannot believe that this can be so. That everything should be explicable in terms of physics (together of course with descriptions of the ways in which the parts are put together—roughly, biology is to physics as radio-engineering is to electro-magnetism) except the occurrence of sensations seems to me to be frankly unbelievable. Such sensations would be 'nomological danglers,' to use Feigl's expression.[5] It is not often realized how odd would be the laws whereby these nomological danglers would dangle. It is sometimes asked, 'Why can't

From *Philosophical Review* 68 (1959): pp. 141–56.

there be psychophysical laws which are of a novel sort, just as the laws of electricity and magnetism were novelties from the standpoint of Newtonian mechanics?'Certainly we are pretty sure in the future to come across new ultimate laws of a novel type, but I expect them to relate simple constituents: for example, whatever ultimate particles are then in vogue. I cannot believe that ultimate laws of nature could relate simple constituents to configurations consisting of perhaps billions of neurons (and goodness knows how many billion billions of ultimate particles) all put together for all the world as though their main purpose in life was to be a negative feedback mechanism of a complicated sort. Such ultimate laws would be like nothing so far known in science. They have a queer 'smell' to them. I am just unable to believe in the nomological danglers themselves, or in the laws whereby they would dangle. If any philosophical arguments seemed to compel us to believe in such things, I would suspect a catch in the argument. In any case it is the object of this paper to show that there are no philosophical arguments which compel us to be dualists.

The above is largely a confession of faith, but it explains why I find Wittgenstein's position (as I construe it) so congenial. For on this view there are, in a sense, no sensations. A man is a vast arrangement of physical particles, but there are not, over and above this, sensations or states of consciousness. There are just behavioral facts about this vast mechanism, such as that it expresses a temptation (behavior disposition) to say 'there is a yellowish-red patch on the wall' or that it goes through a sophisticated sort of wince, that is, says 'I am in pain.' Admittedly Wittgenstein says that though the sensation 'is not a something,' it is nevertheless 'not a nothing either' (paragraph 304), but this need only mean that the word 'ache' has a use. An ache is a thing, but only in the innocuous sense in which the plain man, in the first paragraph of Frege's *Foundations of Arithmetic*, answers the question 'what is the number one?' by 'a thing.' It should be noted that when I assert that to say 'I have a yellowish-orange after-image' is to express a temptation to assert the physical-object statement 'there is a yellowish-orange patch on the wall,' I mean that saying 'I have a yellowish-orange after-image' is (partly) the exercise of the disposition[6] which is the temptation. It is not to *report* that I have the temptation, any more than is 'I love you' normally a report that I love someone. Saying 'I love you' is just part of the behavior which is the exercise of the disposition of loving someone.

Though, for the reasons given above, I am very receptive to the above 'expressive' account of sensation statements, I do not feel that it will quite do the trick. Maybe this is because I have not thought it out sufficiently, but it does seem to me as though, when a person says 'I have an after-image,' he *is* making a genuine report, and that when he says 'I have a pain,' he *is* doing more than 'replace pain-behavior,' and that 'this more' is not just to say that he is in distress. I am not so sure, however, that to admit this is to admit that there are nonphysical correlates of brain processes. Why should not sensations just be brain processes of a certain sort? There are, of course, well-known (as well as lesser-known) philosophical objections to the view that reports of sensations are reports of brain-processes, but I shall try to argue that these arguments are by no means as cogent as is commonly thought to be the case.

Let me first try to state more accurately the thesis that sensations are brain processes. It is not the thesis that, for example, 'after-image' or 'ache' means the same as 'brain process of sort *X*' (where '*X*' is replaced by a description of a certain sort of brain process). It is that, in so far as 'after-image' or 'ache' is a report of a process, it is a report of a process that *happens to be* a brain process. It follows that the thesis does not claim that sensation statements can be *translated* into statements about brain processes.[7] Nor does it claim that the logic of a sensation statement is the same as that of a brain-process statement. All it claims is that in so far as a sensation statement is a report of something, that something is in fact a brain process. Sensations are nothing over and above brain processes. Nations are nothing 'over and above' citizens, but this does not prevent the logic of nation statements being very different from the logic of citizen statements, nor does it insure the translatability of nation statements into citizen statements. (I do not, however, wish to assert that the relation of sensation statements to brain-process statements is very like that of nation statements to citizen statements. Nations do not just *happen to be* nothing over and above citizens, for example. I bring in the 'nations' example merely to make a negative point: that the fact that the logic of A-statements is different from that of B-statements does not insure that A's are anything over and above B's.)

Remarks on Identity

When I say that a sensation is a brain process or that lightning is an electric discharge, I am using 'is' in the sense of strict identity. (Just as in the—in this case necessary—proposition '7 is identical with the smallest prime number greater than 5.') When I say that a sensation is a brain process or that lightning is an electric discharge I do not mean just that the sensation is somehow spatially or temporally continuous with the brain process or that the lightning is just spatially or temporally continuous with the discharge. When on the other hand I say that the successful general is the same person as the small boy who stole the apples I mean only that the successful general I see before me is a time slice[8] of the same four-dimensional object of which the small boy stealing apples is an earlier time slice. However, the four-dimensional object which has the general-I-see-before-me for its late time slice is identical in the strict sense with the four-dimensional object which has the small-boy-stealing-apples for an early time slice. I distinguish these two senses of 'is identical with' because I wish to make it clear that the brain-process doctrine asserts identity in the *strict* sense.

I shall now discuss various possible objections to the view that the processes reported in sensation statements are in fact processes in the brain. Most of us have met some of these objections in our first year as philosophy students. All the more reason to take a good look at them. Others of the objections will be more recondite and subtle.

Objection 1

Any illiterate peasant can talk perfectly well about his after-images, or how things look or feel to him, or about his aches and pains, and yet he may know nothing whatever about neurophysiology. A man may, like Aristotle, believe that the brain is an organ for cooling the body without any impairment of his ability to make true statements about his sensations. Hence the things we are talking about when we describe our sensations cannot be processes in the brain.

Reply

You might as well say that a nation of slugabeds, who never saw the morning star or knew of its existence, or who had never thought of the expression 'the Morning Star,' but who used the expression 'the Evening Star' perfectly well, could not use this expression to refer to the same entity as we refer to (and describe as) 'the Morning Star.'[9]

You may object that the Morning Star is in a sense not the very same thing as the Evening Star, but only something spatiotemporally continuous with it. That is, you may say that the Morning Star is not the Evening Star in the strict sense of 'identity' that I distinguished earlier. I can perhaps forestall this objection by considering the slug-abeds to be New Zealanders and the early risers to be Englishmen. Then the thing the New Zealanders describe as 'the Morning Star' could be the very same thing (in the strict sense) as the Englishmen describe as 'the Evening Star.' And yet they could be ignorant of this fact.

There is, however, a more plausible example. Consider lightning.[10] Modern physical science tells us that lightning is a certain kind of electrical discharge due to ionization of clouds of water-vapor in the atmosphere. This, it is now believed, is what the true nature of lightning is. Note that there are not two things: a flash of lightning and an electrical discharge. There is one thing, a flash of lightning, which is described scientifically as an electrical discharge to the earth from a cloud of ionized water-molecules. The case is not at all like that of explaining a footprint by reference to a burglar. We say that what lightning really is, what its true nature as revealed by science is, is an electric discharge. (It is not the true nature of a footprint to be a burglar.)

To forestall irrelevant objections, I should like to make it clear that by 'lightning' I mean the publicly observable physical object, lightning, not a visual sense-datum of lightning. I say that the publicly observable physical object lightning is in fact the electric discharge, not just a correlate of it. The sense-datum, or at least the having of the sense-datum, the 'look' of lightning, may well in my view be a correlate of the electric discharge. For in my view it is a brain state *caused* by the lightning. But we should no more confuse sensations of lightning with lightning than we confuse sensations of a table with the table.

In short, the reply to Objection 1 is that there can be contingent statements of the form 'A is identical with B,' and a person may well know that something is an A without knowing that it is a B. An illiterate peasant might well be able to talk about his sensations without knowing

about his brain processes, just as he can talk about lightning though he knows nothing of electricity.

Objection 2

It is only a contingent fact (if it is a fact) that when we have a certain kind of sensation there is a certain kind of process in our brain. Indeed it is possible, though perhaps in the highest degree unlikely, that our present physiological theories will be as out of date as the ancient theory connecting mental processes with goings on in the heart. It follows that when we report a sensation we are not reporting a brain-process.

Reply

The objection certainly proves that when we say 'I have an after-image' we cannot *mean* something of the form 'I have such and such a brain process.' But this does not show that what we report (having an after-image) is not *in fact* a brain process. 'I see lightning' does not *mean* 'I see an electric discharge.' Indeed, it is logically possible (though highly unlikely) that the electrical discharge account of lightning might one day be given up. Again, 'I see the Evening Star' does not *mean* the same as 'I see the Morning Star,' and yet 'the Evening Star and the Morning Star are one and the same thing' is a contingent proposition. Possibly Objection 2 derives some of its apparent strength from a 'Fido'—Fido theory of meaning. If the meaning of an expression were what the expression named, then of course it *would* follow from the fact that 'sensation' and 'brain-process' have different meanings that they cannot name one and the same thing.

Objection 3[11]

Even if Objections 1 and 2 do not prove that sensations are something over and above brain-processes, they do prove that the qualities of sensations are something over and above the qualities of brain-processes. That is, it may be possible to get out of asserting the existence of irreducibly psychic processes, but not out of asserting the existence of irreducibly psychic *properties.* For suppose we identify the Morning Star with the Evening Star. Then there must be some properties which logically imply that of being the Morning Star, and quite distinct properties which entail that of being the Evening Star. Again, there must be some properties (for example, that of being a yellow flash) which are logically distinct from those in the physicalist story.

Indeed, it might be thought that the objection succeeds at one jump. For consider the property of 'being a yellow flash.' It might seem that this property lies inevitably outside the physicalist framework within which I am trying to work (either by 'yellow' being an objective emergent property of physical objects, or else by being a power to produce yellow sense-data, where 'yellow,' in this second instantiation of the word, refers to a purely phenomenal or introspectible quality). I must therefore digress for a moment and indicate how I deal with secondary qualities. I shall concentrate on color.

First of all, let me introduce the concept of a normal percipient. One person is more a normal percipient than another if he can make color discriminations that the other cannot. For example, if A can pick a lettuce leaf out of a heap of cabbage leaves, whereas B cannot though he can pick a lettuce leaf out of a heap of beetroot leaves, then A is more normal than B. (I am assuming that A and B are not given time to distinguish the leaves by their slight difference in shape, and so forth.) From the concept of 'more normal than' it is easy to see how we can introduce the concept of 'normal.' Of course, Eskimos may make the finest discriminations at the blue end of the spectrum, Hottentots at the red end. In this case the concept of a normal percipient is a slightly idealized one, rather like that of 'the mean sun' in astronomical chronology. There is no need to go into such subtleties now. I say that 'This is red' means something roughly like 'A normal percipient would not easily pick this out of a clump of geranium petals though he would pick it out of a clump of lettuce leaves.' Of course it does not exactly mean this: a person might know the meaning of 'red' without knowing anything about geraniums, or even about normal percipients. But the point is that a person can be *trained* to say 'This is red' of objects which would not easily be picked out of geranium petals by a normal percipient, and so on. (Note that even a color-blind person can reasonably assert that something is red, though of course he needs to use another human being, not just himself, as his 'color meter.') This account of secondary qualities explains their unimportance in physics. For obviously the discriminations and lack of discriminations made by a very complex neurophysiological mechanism are hardly likely to

correspond to simple and nonarbitrary distinctions in nature.

I therefore elucidate colors as powers, in Locke's sense, to evoke certain sorts of discriminatory responses in human beings. They are also, of course, powers to cause sensations in human beings (an account still nearer Locke's). But these sensations, I am arguing, are identifiable with brain processes.

Now how do I get over the objection that a sensation can be identified with a brain process only if it has some phenomenal property, not possessed by brain processes, whereby one-half of the identification may be, so to speak, pinned down?

My suggestion is as follows. When a person says, 'I see a yellowish-orange after-image,' he is saying something like this: '*There is something going on which is like what is going on when* I have my eyes open, am awake, and there is an orange illuminated in good light in front of me, that is, when I really see an orange.' (And there is no reason why a person should not say the same thing when he is having a veridical sense-datum, so long as we construe 'like' in the last sentence in such a sense that something can be like itself.) Notice that the italicized words, namely 'there is something going on which is like what is going on when,' are all quasi-logical or topic-neutral words. This explains why the ancient Greek peasant's reports about his sensations can be neutral between dualistic metaphysics or my materialistic metaphysics. It explains how sensations can be brain-processes and yet how those who report them need know nothing about brain-processes. For he reports them only very abstractly as 'something going on which is like what is going on when . . .' Similarly, a person may say 'someone is in the room,' thus reporting truly that the doctor is in the room, even though he has never heard of doctors. (There are not two people in the room: 'someone'*and* the doctor.) This account of sensation statements also explains the singular elusiveness of 'raw feels'—why no one seems to be able to pin any properties on them.[12] Raw feels, in my view, are colorless for the very same reason that *something* is colorless. This does not mean that sensations do not have properties, for if they are brain-processes they certainly have properties. It only means that in speaking of them as being like or unlike one another we need not know or mention these properties.

This, then, is how I would reply to Objection 3. The strength of my reply depends on the possibility of our being able to report that one thing is like another without being able to state the respect in which it is like. I am not sure whether this is so or not, and that is why I regard Objection 3 as the strongest with which I have to deal.

Objection 4

The after-image is not in physical space. The brain-process is. So the after-image is not a brain-process.

Reply

This is an *ignoratio elenchi.* I am not arguing that the after-image is a brain-process, but that the experience of having an after-image is a brain-process. It is the *experience* which is reported in the introspective report. Similarly, if it is objected that the after-image is yellowy-orange but that a surgeon looking into your brain would see nothing yellowy-orange, my reply is that it is the experience of seeing yellowy-orange that is being described, and this experience is not a yellowy-orange something. So to say that a brain-process cannot be yellowy-orange is not to say that a brain-process cannot in fact be the experience of having a yellowy-orange after-image. There is, in a sense, no such thing as an after image or a sense-datum, though there is such a thing as the experience of having an image, and this experience is described indirectly in material object language, not in phenomenal language, for there is no such thing.[13] We describe the experience by saying, in effect, that it is like the experience we have when, for example, we really see a yellowy-orange patch on the wall. Trees and wallpaper can be green, but not the experience of seeing or imagining a tree or wallpaper. (Or if they are described as green or yellow this can only be in a derived sense.)

Objection 5

It would make sense to say of a molecular movement in the brain that it is swift or slow, straight or circular, but it makes no sense to say this of the experience of seeing something yellow.

Reply

So far we have not given sense to talk of experiences as swift or slow, straight or circular. But I am not claiming that 'experience' and

'brain-process' mean the same or even that they have the same logic. 'Somebody' and 'the doctor' do not have the same logic, but this does not lead us to suppose that talking about somebody telephoning is talking about someone over and above, say, the doctor. The ordinary man when he reports an experience is reporting that something is going on, but he leaves it open as to what sort of thing is going on, whether in a material solid medium, or perhaps in some sort of gaseous medium, or even perhaps in some sort of non-spatial medium (if this makes sense). All that I am saying is that 'experience' and 'brain-process' may in fact refer to the same thing, and if so we may easily adopt a convention (which is not a change in our present rules for the use of experience words but an addition to them) whereby it would make sense to talk of an experience in terms appropriate to physical processes.

Objection 6

Sensations are private, brain processes are *public*. If I sincerely say, 'I see a yellowish-orange after-image' and I am not making a verbal mistake, then I cannot be wrong. But I can be wrong about a brain-process. The scientist looking into my brain might be having an illusion. Moreover, it makes sense to say that two or more people are observing the same brain-process but not that two or more people are reporting the same inner experience.

Reply

This shows that the language of introspective reports has a different logic from the language of material processes. It is obvious that until the brain-process theory is much improved and widely accepted there will be no *criteria* for saying 'Smith has an experience of such-and-such a sort' *except* Smith's introspective reports. So we have adopted a rule of language that (normally) what Smith says goes.

Objection 7

I can imagine myself turned to stone and yet having images, aches, pains, and so on.

Reply

I can imagine that the electrical theory of lightning is false, that lightning is some sort of purely optical phenomenon. I can imagine that lightning is not an electrical discharge. I can imagine that the Evening Star is not the Morning Star. But it is. All the objection shows is that 'experience' and 'brain-process' do not have the same meaning. It does not show that an experience is not in fact a brain process.

This objection is perhaps much the same as one which can be summed up by the slogan: 'What can be composed of nothing cannot be composed of anything.'[14] The argument goes as follows: on the brain-process thesis the identity between the brain-process and the experience is a contingent one. So it is logically possible that there should be no brain-process, and no process of any other sort, either (no heart process, no kidney process, no liver process). There would be the experience but no 'corresponding' physiological process with which we might be able to identify it empirically.

I suspect that the objector is thinking of the experience as a ghostly entity. So it is composed of something, not of nothing, after all. On his view it is composed of ghost stuff, and on mine it is composed of brain stuff. Perhaps the counter-reply will be[15] that the experience is simple and uncompounded, and so it is not composed of anything after all. This seems to be a quibble, for, if it were taken seriously, the remark 'What can be composed of nothing cannot be composed of anything' could be recast as an a priori argument against Democritus and atomism and for Descartes and infinite divisibility. And it seems odd that a question of this sort could be settled a priori. We must therefore construe the word 'composed' in a very weak sense, which would allow us to say that even an indivisible atom is composed of something (namely, itself). The dualist cannot really say that an experience can be composed of nothing. For he holds that experiences are something over and above material processes, that is, that they are a sort of ghost stuff. (Or perhaps ripples in an underlying ghost stuff.) I say that the dualist's hypothesis is a perfectly intelligible one. But I say that experiences are not to be identified with ghost stuff but with brain stuff. This is another hypothesis, and in my view a very plausible one. The present argument cannot knock it down a priori.

Objection 8

The 'beetle in the box' objection (see Wittgenstein, *Philosophical Investigations*, paragraph 293). How could descriptions of

experiences, if these are genuine reports, get a foothold in language? For any rule of language must have public criteria for its correct application.

Reply

The change from describing how things are to describing how we feel is just a change from uninhibitedly saying 'this is so' to saying 'this looks so.' That is, when the naive person might be tempted to say, 'There is a patch of light on the wall which moves whenever I move my eyes' or 'A pin is being stuck into me,' we have learned how to resist this temptation and say 'It *looks as though* there is a patch of light on the wallpaper' or 'It *feels as though* someone were sticking a pin into me.' The introspective account tells us about the individual's state of consciousness in the same way as does 'I see a patch of light' or 'I feel a pin being stuck into me': it differs from the corresponding perception statement in so far as (a) in the perception statement the individual 'goes beyond the evidence of his senses' in describing his environment and (b) in the introspective report he withholds descriptive epithets he is inclined to ascribe to the environment, perhaps because he suspects that they may not be appropriate to the actual state of affairs. Psychologically speaking, the change from talking about the environment to talking about one's state of consciousness is simply a matter of inhibiting descriptive reactions not justified by appearances alone, and of disinhibiting descriptive reactions which are normally inhibited because the individual has learned that they are unlikely to provide a reliable guide to the state of the environment in the prevailing circumstances.[16] To say that something looks green to me is to say that my experience is like the experience I get when I see something that really is green. In my reply to Objection 3, I pointed out the extreme openness or generality of statements which report experiences. This explains why there is no language of private qualities. (Just as 'someone,' unlike 'the doctor,' is a colorless word.)[17]

If it is asked what is the difference between those brain processes which, in my view, are experiences and those brain processes which are not, I can only reply that this is at present unknown. But it does not seem to me altogether fanciful to conjecture that the difference may in part be that between perception and reception (in Dr. D. M. MacKay's terminology) and that the type of brain process which is an experience might be identifiable with MacKay's active 'matching response.'[18]

I have now considered a number of objections to the brain-process thesis. I wish now to conclude by some remarks on the logical status of the thesis itself. U. T. Place seems to hold that it is a straight-out scientific hypothesis.[19] If so, he is partly right and partly wrong. If the issue is between (say) a brain-process thesis and a heart thesis, or a liver thesis, or a kidney thesis, then the issue is a purely empirical one, and the verdict is overwhelmingly in favor of the brain. The right sorts of things don't go on in the heart, liver, or kidney, nor do these organs possess the right sort of complexity of structure. On the other hand, if the issue is between a brain-or-heart-or-liver-or-kidney thesis (that is, some form of materialism) on the one hand and epiphenomenalism on the other hand, then the issue is not an empirical one. For there is no conceivable experiment which could decide between materialism and epiphenomenalism. This latter issue is not like the average straight-out empirical issue in science, but like the issue between the nineteenth-century English naturalist Philip Gosse[20] and the orthodox geologists and paleontologists of his day. According to Gosse, the earth was created about 4000 B.C. exactly as described in Genesis, with twisted rock strata, 'evidence' of erosion, and so forth, and all sorts of fossils, all in their appropriate strata, just as if the usual evolutionist story had been true. Clearly this theory is in a sense irrefutable: no evidence can possibly tell against it. Let us ignore the theological setting in which Philip Gosse's hypothesis had been placed, thus ruling out objections of a theological kind, such as 'what a queer God who would go to such elaborate lengths to deceive us.' Let us suppose that it is held that the universe just *began* in 4004 B.C. with the initial conditions just everywhere as they were in 4004 B.C., and in particular that our own planet began with sediment in the rivers, eroded cliffs, fossils in the rocks, and so on. No scientist would ever entertain this as a serious hypothesis, consistent though it is with all possible evidence. The hypothesis offends against the principles of parsimony and simplicity. There would be far too many brute and inexplicable facts. Why are pterodactyl bones just as they are? No explanation in terms of the evolution of pterodactyls from earlier forms of life would any longer be possible. We would have millions of facts about the world as it was in 4004 B.C. that just have to be *accepted.*

The issue between the brain-process theory and epiphenomenalism seems to be of the above sort. (Assuming that a behavioristic reduction of introspective reports is not possible.) If it be agreed that there are no cogent philosophical arguments which force us into accepting dualism, and if the brain process theory and dualism are equally consistent with the facts, then the principles of parsimony and simplicity seem to me to decide overwhelmingly in favor of the brain-process theory. As I pointed out earlier, dualism involves a large number of irreducible psychophysical laws (whereby the 'nomological danglers' dangle) of a queer sort, that just have to be taken on trust, and are just as difficult to swallow as the irreducible facts about the paleontology of the earth with which we are faced on Philip Gosse's theory.

NOTES

1. This paper takes its departure from arguments to be found in U. T. Place's "Is Consciousness a Brain Process?" (*British Journal of Psychology*, XLVII, 1956, pp. 44–50). I have had the benefit of discussing Place's thesis in a good many universities in the United States and Australia, and I hope that the present paper answers objections to his thesis which Place has not considered, and presents his thesis in a more nearly unobjectionable form. This paper is meant also to supplement "The 'Mental' and the 'Physical,'" by H. Feigl (in *Minnesota Studies in the Philosophy of Science*, II, pp. 370–497), which argues for much the same thesis as Place's.
2. Some philosophers of my acquaintance, who have the advantage over me in having known Wittgenstein, would say that this interpretation of him is too behavioristic. However, it seems to me a very natural interpretation of his printed words, and whether or not it is Wittgenstein's real view it is certainly an interesting and important one. I wish to consider it here as a possible rival both to the 'brain-process' thesis and to straight-out old-fashioned dualism.
3. See Ryle, *Concept of Mind.* (New York: Hutchinson, 1949), p. 93.
4. On this point see Paul Oppenheim and Hilary Putnam, "Unity of Science as a Working Hypothesis," in *Minnesota Studies in the Philosophy of Science* II, pp. 3–36; also my note "Plausible Reasoning in Philosophy," *Mind* LXVI (1957): pp. 75–78.
5. Feigl, op. cit., p. 428.
6. Wittgenstein did not like the word 'disposition.' I am using it to put in a nutshell (and perhaps inaccurately) the view which I am attributing to Wittgenstein. I should like to repeat that I do not wish to claim that my interpretation of Wittgenstein is correct. Some of those who knew him do not interpret him in this way. It is merely a view which I find myself extracting from his printed words and which I think is important and worth discussing for its own sake.
7. See Place, op. cit., p. 45, near top, and Feigl, op. cit., p. 390, near top.
8. See J. H. Woodger, *"Theory Construction"* in *International Encyclopedia of Unified Science*, II, number 5, (Chicago, 1939), p. 38. I here permit myself to speak loosely. For warnings against possible ways of going wrong with this sort of talk, see my note "Spatialising Time," *Mind* LXIV (1955): pp. 239–41.
9. Cf. Feigl, op. cit., p. 439.
10. See Place, op. cit., p. 47; also Feigl, op. cit. p. 438.
11. I think this objection was first put to me by Professor Max Black. I think it is the most subtle of any of those I have considered, and the one which I am least confident of having satisfactorily met.
12. See B. A. Farrell, "Experience," *Mind* LIX (1950), especially p. 174.
13. Dr. J. R. Smythies claims that a sense-datum language could be taught independently of the material object language ("A Note on the Fallacy of the 'Phenomenological Fallacy,'" *British Journal of Psychology* XLVIII [1957]: pp. 141–44). I am not so sure of this: there must be some public criteria for a person having got a rule wrong before we can teach him the rule. I suppose someone might *accidentally* learn color words by Dr. Smythies' procedure. I am not, of course, denying that we can learn a sense-datum language in the sense that we can learn to report our experience. Nor would Place deny it.
14. I owe this objection to Mr. C. B. Martin. I gather that he no longer wishes to maintain this objection, at any rate in its present form.
15. Martin did not make this reply, but one of his students did.
16. I owe this point to Place, in correspondence.
17. The 'beetle in the box' objection is, *if it is sound*, an objection to *any* view, and in particular the Cartesian one, that introspective reports are genuine reports. So it is no objection to a weaker thesis that I would be concerned to uphold, namely, that if introspective reports of 'experiences' are genuinely reports, then the things they are reports of are in fact brain processes.
18. See his article "Towards an Information-Flow Model of Human Behaviour," *British Journal of Psychology* XLVII (1956): pp. 30–43.
19. Op. cit.
20. See the entertaining account of Gosse's book, *Omphalos*, by Martin Gardner in *Fads and Fallacies in the Name of Science*, 2nd ed. (New York: Dover Publications, 1957).

The Nature of Mental States

Hilary Putnam

The typical concerns of the Philosopher of Mind might be represented by three questions: (1) How do we know that other people have pains? (2) Are pains brain states? (3) What is the analysis of the concept *pain?* I do not wish to discuss questions (1) and (3) in this paper. I shall say something about question (2).[1]

I. Identity Questions

'Is pain a brain state?' (Or, 'Is the property of having a pain at time *t* a brain state?')[2] It is impossible to discuss this question sensibly without saying something about the peculiar rules which have grown up in the course of the development of 'analytical philosophy'—rules which, far from leading to an end to all conceptual confusions, themselves represent considerable conceptual confusion. These rules—which are, of course, implicit rather than explicit in the practice of most analytical philosophers—are (1) that a statement of the form 'being *A* is being *B*' (e.g., 'being in pain is being in a certain brain state') can be *correct* only if it follows, in some sense, from the meaning of the terms *A* and *B;* and (2) that a statement of the form 'being *A* is being *B*' can be philosophically *informative* only if it is in some sense reductive (e.g. 'being in pain is having a certain unpleasant sensation' is not philosophically informative; 'being in pain is having a certain behavior disposition' is, if true, philosophically informative). These rules are excellent rules if we still believe that the program of reductive analysis (in the style of the 1930s) can be carried out; if we don't, then they turn analytical philosophy into a mug's game, at least so far as 'is' questions are concerned.

In this paper I shall use the term 'property' as a blanket term for such things as being in pain, being in a particular brain state, having a particular behavior disposition, and also for magnitudes such as temperature, etc.—i.e., for things which can naturally be represented by one-or-more-place predicates or functors. I shall use the term 'concept' for things which can be identified with synonymy-classes of expressions. Thus the concept *temperature* can be identified (I maintain) with the synonymy-class of the word 'temperature.'[3] (This is like saying that the number 2 can be identified with the class of all pairs. This is quite a different statement from the peculiar statement that 2 *is* the class of all pairs. I do not maintain that concepts *are* synonymy-classes, whatever that might mean, but that they can be identified with synonymy classes, for the purpose of formalization of the relevant discourse.)

The question 'What is the concept *temperature?*' is a very 'funny' one. One might take it to mean 'What is temperature? Please take my question as a conceptual one.' In that case an answer might be (pretend for a moment 'heat' and 'temperature' are synonyms) 'temperature is heat,' or even 'the concept of temperature is the same concept as the concept of heat.' Or one might take it to mean 'What are *concepts*, really? For example, what is 'the concept of temperature'?' In that case heaven knows what an 'answer' would be. (Perhaps it would be the statement that concepts *can be identified with* synonymy-classes.)

Of course, the question 'What is the property temperature?' is also 'funny.' And one way of interpreting it is to take it as a question about the concept of temperature. But this is not the way a physicist would take it.

The effect of saying that the property P_1 can be identical with the property P_2 only if the terms P_1, P_2 are in some suitable sense 'synonyms' is, to all intents and purposes, to collapse the two notions of 'property' and 'concept' into a single notion. The view that concepts (intensions) *are* the same as properties has been explicitly advocated by Carnap (e.g., in *Meaning and Necessity*). This seems an unfortunate view, since 'temperature is mean molecular kinetic energy' appears to be a perfectly good example of a true statement of identity of

Originally published as "Psychological Predicates" in *Art, Mind, and Religion*, W. H. Capitan and D. D. Merrill, eds.,(University of Pittsburgh Press, 1973), pp. 37–48. Reprinted with permission of University of Pittsburgh Press.

properties, whereas 'the concept of temperature is the same concept as the concept of mean molecular kinetic energy' is simply false.

Many philosophers believe that the statement 'pain is a brain state' violates some rules or norms of English. But the arguments offered are hardly convincing. For example, if the fact that I can know that I am in pain without knowing that I am in brain state *S* shows that pain cannot be brain state *S*, then, by exactly the same argument, the fact that I can know that the stove is hot without knowing that the mean molecular kinetic energy is high (or even that molecules exist) shows that it is *false* that temperature is mean molecular kinetic energy, physics to the contrary. In fact, all that immediately follows from the fact that I can know that I am in pain without knowing that I am in brain state *S* is that the concept of pain is not the same concept as the concept of being in brain state *S*. But either pain, or the state of being in pain, or some pain, or some pain state, might still be brain state *S*. After all, the concept of temperature is not the same concept as the concept of mean molecular kinetic energy. But temperature is mean molecular kinetic energy.

Some philosophers maintain that both 'pain is a brain state' and 'pain states are brain states' are unintelligible. The answer is to explain to these philosophers, as well as we can, given the vagueness of all scientific methodology, what sorts of considerations lead one to make an empirical reduction (i.e. to say such things as 'water is H_2O,' 'light is electromagnetic radiation,' 'temperature is mean molecular kinetic energy'). If, without giving reasons, he still maintains in the face of such examples that one cannot imagine parallel circumstances for the use of 'pains are brain states' (or, perhaps, 'pain states are brain states') one has grounds to regard him as perverse.

Some philosophers maintain that 'P_1 is P_2' is something that can be true, when the 'is' involved is the 'is' of empirical reduction, only when the properties P_1 and P_2 are (a) associated with a spatio-temporal region; and (b) the region is one and the same in both cases. Thus 'temperature is mean molecular kinetic energy' is an admissible empirical reduction, since the temperature and the molecular energy are associated with the same space-time region, but 'having a pain in my arm is being in a brain state' is not, since the spatial regions involved are different.

This argument does not appear very strong. Surely no one is going to be deterred from saying that mirror images are light reflected from an object and then from the surface of a mirror by the fact that an image can be 'located' three feet *behind* the mirror! (Moreover, one can always find *some* common property of the reductions one is willing to allow—e.g., temperature is mean molecular kinetic energy—which is not a property of some one identification one wishes to disallow. This is not very impressive unless one has an argument to show that the very purposes of such identification depend upon the common property in question.)

Again, other philosophers have contended that all the predictions that can be derived from the conjunction of neurophysiological laws with such statements as 'pain states are such-and-such brain states' can equally well be derived from the conjunction of the same neurophysiological laws with 'being in pain is correlated with such-and-such brain states,' and hence (sic!) there can be no methodological grounds for saying that pains (or pain states) *are* brain states, as opposed to saying that they are *correlated* (invariantly) with brain states. This argument, too, would show that light is only correlated with electromagnetic radiation. The mistake is in ignoring the fact that, although the theories in question may indeed lead to the same predictions, they open and exclude different *questions*. 'Light is invariantly correlated with electromagnetic radiation' would leave open the questions 'What is the light then, if it isn't the same as the electromagnetic radiation?' and 'What makes the light accompany the electromagnetic radiation?'—questions which are excluded by saying that the light *is* the electromagnetic radiation. Similarly, the purpose of saying that pains are brain states is precisely to exclude from empirical meaningfulness the questions 'What is the pain, then, if it isn't the same as the brain state?' and 'What makes the pain accompany the brain state?' If there are grounds to suggest that these questions represent, so to speak, the wrong way to look at the matter, then those grounds are grounds for a theoretical identification of pains with brain states.

If all arguments to the contrary are unconvincing, shall we then conclude that it is meaningful (and perhaps true) to say either that pains are brain states or that pain states are brain states?

1. It is perfectly meaningful (violates no 'rule of English,' involves no 'extension of usage') to say 'pains are brain states.'
2. It is not meaningful (involves a 'changing of meaning' or 'an extension of usage,' etc.) to say 'pains are brain states.'

My own position is not expressed by either (1) or (2). It seems to me that the notions 'change of meaning' and 'extension of usage' are simply so ill-defined that one cannot in fact say *either* (1) or (2). I see no reason to believe that either the linguist, or the man-on-the-street, or the philosopher possesses today a notion of 'change of meaning' applicable to such cases as the one we have been discussing. The *job* for which the notion of change of meaning was developed in the history of the language was just a *much* cruder job than this one.

But, if we don't assert either (1) or (2)—in other words, if we regard the 'change of meaning' issue as a pseudo-issue in this case—then how are we to discuss the question with which we started? 'Is pain a brain state?'

The answer is to allow statements of the form 'pain is *A*,' where 'pain' and '*A*' are in no sense synonyms, and to see whether any such statement can be found which might be acceptable on empirical and methodological grounds. This is what we shall now proceed to do.

II. Is Pain a Brain State?

We shall discuss 'Is pain a brain state?' then. And we have agreed to waive the 'change of meaning' issue.

Since I am discussing not what the concept of pain comes to, but what pain is, in a sense of 'is' which requires empirical theory-construction (or, at least, empirical speculation), I shall not apologize for advancing an empirical hypothesis. Indeed, my strategy will be to argue that pain is *not* a brain state, not on a priori grounds, but on the grounds that another hypothesis is more plausible. The detailed development and verification of my hypothesis would be just as Utopian a task as the detailed development and verification of the brain-state hypothesis. But the putting-forward, not of detailed and scientifically 'finished' hypotheses, but of schemata for hypotheses, has long been a function of philosophy. I shall, in short, argue that pain is not a brain state, in the sense of a physical-chemical state of the brain (or even the whole nervous system), but another *kind* of state entirely. I propose the hypothesis that pain, or the state of being in pain, is a functional state of a whole organism.

To explain this it is necessary to introduce some technical notions. In previous papers I have explained the notion of a Turing Machine and discussed the use of this notion as a model for an organism. The notion of a Probabilistic Automaton is defined similarly to a Turing Machine, except that the transitions between 'states' are allowed to be with various probabilities rather than being 'deterministic.' (Of course, a Turing Machine is simply a special kind of Probabilistic Automaton, one with transition probabilities 0, 1.) I shall assume the notion of a Probabilistic Automaton has been generalized to allow for 'sensory inputs' and 'motor outputs'—that is, the Machine Table specifies, for every possible combination of a 'state' and a complete set of 'sensory inputs,' an 'instruction' which determines the probability of the next 'state,' and also the probabilities of the 'motor outputs.' (This replaces the idea of the Machine as printing on a tape.) I shall also assume that the physical realization of the sense organs responsible for the various inputs, and of the motor organs, is specified, but that the 'states' and the 'inputs' themselves are, as usual, specified only 'implicitly'—i.e., by the set of transition probabilities given by the Machine Table.

Since an empirically given system can simultaneously be a 'physical realization' of many different Probabilistic Automata, I introduce the notion of a *Description* of a system. A Description of *S* where *S* is a system, is any true statement to the effect that *S* possesses distinct states $S_1, S_2 \ldots, S_n$ which are related to one another and to the motor outputs and sensory inputs by the transition probabilities given in such-and-such a Machine Table. The Machine Table mentioned in the Description will then be called the Functional Organization of *S* relative to that Description, and the S_i such that *S* is in state S_i at a given time will be called the Total State of *S* (at that time) relative to that Description. It should be noted that knowing the Total State of a system relative to a Description involves knowing a good deal about how the system is likely to 'behave,' given various combinations of sensory inputs, but does *not* involve knowing the physical realization of the S_i as, e.g., physical-chemical states of the brain. The S_i, to repeat, are specified only *implicitly* by the Description—i.e., specified *only* by the set of transition probabilities given in the Machine Table.

The hypothesis that 'being in pain is a functional state of the organism' may now be spelled out more exactly as follows:

1. All organisms capable of feeling pain are Probabilistic Automata.
2. Every organism capable of feeling pain possesses at least one Description of a

certain kind (i.e., being capable of feeling pain *is* possessing an appropriate kind of Functional Organization).

3. No organism capable of feeling pain possesses a decomposition into parts which separately possess Descriptions of the kind referred to in (2).
4. For every Description of the kind referred to in (2), there exists a subset of the sensory inputs such that an organism with that Description is in pain when and only when some of its sensory inputs are in that subset.

This hypothesis is admittedly vague, though surely no vaguer than the brain-state hypothesis in its present form. For example, one would like to know more about the kind of Functional Organization that an organism must have to be capable of feeling pain, and more about the marks that distinguish the subset of the sensory inputs referred to in (4). With respect to the first question, one can probably say that the Functional Organization must include something that resembles a 'preference function,' or at least a preference partial ordering, and something that resembles an 'inductive logic' (i.e., the Machine must be able to 'learn from experience'). (The meaning of these conditions, for Automata models, is discussed in my paper 'The Mental Life of Some Machines.') In addition, it seems natural to require that the Machine possess 'pain sensors,' i.e., sensory organs which normally signal damage to the Machine's body, or dangerous temperatures, pressures, etc., which transmit a special subset of the inputs, the subset referred to in (4). Finally, and with respect to the second question, we would want to require at least that the inputs in the distinguished subset have a high disvalue on the Machine's preference function or ordering (further conditions are discussed in 'The Mental Life of Some Machines'). The purpose of condition (3) is to rule out such 'organisms' (if they can count as such) as swarms of bees as single pain-feelers. The condition (1) is, obviously, redundant, and is only introduced for expository reasons. (It is, in fact, empty, since everything is a Probabilistic Automaton under *some* Description.)

I contend, in passing, that this hypothesis, in spite of its admitted vagueness, is far *less* vague than the 'physical-chemical state' hypothesis is today, and far more susceptible to investigation of both a mathematical and an empirical kind. Indeed, to investigate this hypothesis is just to attempt to produce 'mechanical' models of organisms—and isn't this, in a sense, just what psychology is about? The difficult step, of course, will be to pass from models of *specific* organisms to a *normal form* for the psychological description of organisms—for this is what is required to make (2) and (4) precise. But this too seems to be an inevitable part of the program of psychology.

I shall now compare the hypothesis just advanced with (a) the hypothesis that pain is a brain state, and (b) the hypothesis that pain is a behavior disposition.

III. Functional State versus Brain State

It may, perhaps, be asked if I am not somewhat unfair in taking the brain-state theorist to be talking about *physical-chemical* states of the brain. But (a) these are the only sorts of states ever mentioned by brain-state theorists. (b) The brain-state theorist usually mentions (with a certain pride, slightly reminiscent of the Village Atheist) the incompatibility of his hypothesis with all forms of dualism and mentalism. This is natural if physical-chemical states of the brain are what is at issue. However, functional states of whole systems are something quite different. In particular, the functional-state hypothesis is *not* incompatible with dualism! Although it goes without saying that the hypothesis is 'mechanistic' in its inspiration, it is a slightly remarkable fact that a system consisting of a body and a 'soul,' if such things there be, can perfectly well be a Probabilistic Automaton. (c) One argument advanced by Smart is that the brain-state theory assumes only 'physical' properties, and Smart finds 'non-physical' properties unintelligible. The Total States and the 'inputs' defined above are, of course, neither mental nor physical per se, and I cannot imagine a functionalist advancing this argument. (d) If the brain-state theorist does mean (or at least allow) states other than physical-chemical states, then his hypothesis is completely empty, at least until he specifies *what* sort of 'states' he *does* mean.

Taking the brain-state hypothesis in this way, then, what reasons are there to prefer the functional-state hypothesis over the brain-state hypothesis? Consider what the brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that *any* organism (not just a mammal) is in pain if and only if (a) it possesses a brain of a suitable

physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are mollusca, and certainly feel pain), etc. At the same time, it must *not* be a possible (physically possible) state of the brain of any physically possible creature that cannot feel pain. Even if such a state can be found, it must be nomologically certain that it will also be a state of the brain of any extra-terrestrial life that may be found that will be capable of feeling pain before we can even entertain the supposition that it may *be* pain.

It is not altogether impossible that such a state will be found. Even though octopus and mammal are examples of parallel (rather than sequential) evolution, for example, virtually identical structures (physically speaking) have evolved in the eye of the octopus and in the eye of the mammal, notwithstanding the fact that this organ has evolved from different kinds of cells in the two cases. Thus it is at least possible that parallel evolution, all over the universe, might *always* lead to *one and the same* physical 'correlate' of pain. But this is certainly an ambitious hypothesis.

Finally, the hypothesis becomes still more ambitious when we realize that the brain-state theorist is not just saying that *pain* is a brain state; he is, of course, concerned to maintain that *every* psychological state is a brain state. Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say 'hungry'), but whose physical-chemical 'correlate' is different in the two cases, the brain-state theory has collapsed. It seems to me overwhelmingly probable that we can do this. Granted, in such a case the brain-state theorist can save himself by ad hoc assumptions (e.g., defining the disjunction of two states to be a single 'physical-chemical state'), but this does not have to be taken seriously.

Turning now to the considerations *for* the functional-state theory, let us begin with the fact that we identify organisms as in pain, or hungry, or angry, or in heat, etc., on the basis of their *behavior.* But it is a truism that similarities in the behavior of two systems are at least a reason to suspect similarities in the functional organization of the two systems, and a much *weaker* reason to suspect similarities in the actual physical details. Moreover, we expect the various psychological states—at least the basic ones, such as hunger, thirst, aggression, etc.—to have more or less similar 'transition probabilities' (within wide and ill-defined limits, to be sure) with each other and with behavior in the case of different species, because this is an artifact of the way in which we identify these states. Thus, we would not count an animal as *thirsty* if its 'unsatiated' behavior did not seem to be directed toward drinking and was not followed by 'satiation for liquid.' Thus any animal that we count as capable of these various states will at least *seem* to have a certain rough kind of functional organization. And, as already remarked, if the program of finding psychological laws that are not species-specific—i.e., of finding a normal form for psychological theories of different species—ever succeeds, then it will bring in its wake a delineation of the kind of functional organization that is necessary and sufficient for a given psychological state, as well as a precise definition of the notion 'psychological state.' In contrast, the brain-state theorist has to hope for the eventual development of neurophysiological laws that are species-independent, which seems much less reasonable than the hope that psychological laws (of a sufficiently general kind) may be species-independent, or, still weaker, that a species-independent *form* can be found in which psychological laws can be written.

IV. Functional State versus Behavior Disposition

The theory that being in pain is neither a brain state nor a functional state but a behavior disposition has one apparent advantage: it appears to agree with the way in which we verify that organisms are in pain. We do not in practice know anything about the brain state of an animal when we say that it is in pain; and we possess little if any knowledge of its functional organization, except in a crude intuitive way. In fact, however, this 'advantage' is no advantage at all: for, although statements about how we verify that x is A may have a good deal to do with what the concept of being A comes to, they have precious little to do with what the property A is. To argue on the ground just mentioned that pain is neither a brain state nor a functional state is like arguing that heat is not mean molecular kinetic energy from the fact that ordinary people do not (they think) ascertain the mean molecular kinetic energy of something when they verify that it is hot or cold. It is not necessary that they

should; what is necessary is that the marks that they take as indications of heat should in fact be explained by the mean molecular kinetic energy. And, similarly, it is necessary to our hypothesis that the marks that are taken as behavioral indications of pain should be explained by the fact that the organism is in a functional state of the appropriate kind, but not that speakers should *know* that this is so.

The difficulties with 'behavior disposition' accounts are so well known that I shall do little more than recall them here. The difficulty—it appears to be more than 'difficulty,' in fact—of specifying the required behavior disposition except as 'the disposition of *X* to behave as if *X* were in *pain*,' is the chief one, of course. In contrast, we *can* specify the functional state with which we propose to identify pain, at least roughly, without using the notion of pain. Namely, the functional state we have in mind is the state of receiving sensory inputs which play a certain role in the Functional Organization of the organism. This role is characterized, at least partially, by the fact that the sense organs responsible for the inputs in question are organs whose function is to detect damage to the body, or dangerous extremes of temperature, pressure, etc., and by the fact that the 'inputs' themselves, whatever their physical realization, represent a condition that the organism assigns a high disvalue to. As I stressed in 'The Mental Life of Some Machines,' this does *not* mean that the Machine will always *avoid* being in the condition in question ('pain'); it only means that the condition will be avoided unless not avoiding it is necessary to the attainment of some more highly valued goal. Since the behavior of the Machine (in this case, an organism) will depend not merely on the sensory inputs, but also on the Total State (i.e., on other values, beliefs, etc.), it seems hopeless to make any general statement about how an organism in such a condition *must* behave; but this does not mean that we must abandon hope of characterizing the condition. Indeed, we have just characterized it.[4]

Not only does the behavior-disposition theory seem hopelessly vague; if the 'behavior' referred to is peripheral behavior, and the relevant stimuli are peripheral stimuli (e.g., we do not say anything about what the organism will do if its brain is operated upon), then the theory seems clearly false. For example, two animals with all motor nerves cut will have the same actual and potential 'behavior' (viz., none to speak of); but if one has cut pain fibers and the other has uncut pain fibers, then one will feel pain and the other won't. Again, if one person has cut pain fibers, and another suppresses all pain responses deliberately due to some strong compulsion, then the actual and potential peripheral behavior may be the same, but one will feel pain and the other won't. (Some philosophers maintain that this last case is conceptually impossible, but the only evidence for this appears to be that *they* can't, or don't want to, conceive of it.)[5] If, instead of pain, we take some sensation the 'bodily expression' of which is easier to suppress—say, a slight coolness in one's left little finger—the case becomes even clearer.

Finally, even if there *were* some behavior disposition invariantly correlated with pain (species-independently!), and specifiable without using the term 'pain,' it would still be more plausible to identify being in pain with some state whose presence *explains* this behavior disposition—the brain state or functional state—than with the behavior disposition itself. Such considerations of plausibility may be somewhat subjective; but if other things *were* equal (of course, they aren't) why shouldn't we allow considerations of plausibility to play the deciding role?

V. Methodological Considerations

So far we have considered only what might be called the 'empirical' reasons for saying that being in pain is a functional state, rather than a brain state or a behavior disposition; viz., that it seems more likely that the functional state we described is invariantly 'correlated' with pain, species-independently, than that there is either a physical-chemical state of the brain (must an organism have a *brain* to feel pain? perhaps some ganglia will do) or a behavior disposition so correlated. If this is correct, then it follows that the identification we proposed is at least a candidate for consideration. What of methodological considerations?

The methodological considerations are roughly similar in all cases of reduction, so no surprises need be expected here. First, identification of psychological states with functional states means that the laws of psychology can be derived from statements of the form 'such-and-such organisms have such-and-such Descriptions' together with the identification statements ('being in pain is such-and-such a functional state,' etc.). Secondly, the presence

of the functional state (i.e., of inputs which play the role we have described in the Functional Organization of the organism) is not merely 'correlated with' but actually explains the pain behavior on the part of the organism. Thirdly, the identification serves to exclude questions which (if a naturalistic view is correct) represent an altogether wrong way of looking at the matter, e.g., 'What *is* pain if it isn't either the brain state or the functional state?' and 'What causes the pain to be always accompanied by this sort of functional state?' In short, the identification is to be tentatively accepted as a theory which leads to both fruitful predictions and to fruitful *questions*, and which serves to discourage fruitless and empirically senseless questions, where by 'empirically senseless' I mean 'senseless' not merely from the standpoint of verification, but from the standpoint of what there in fact *is*.

NOTES

1. I have discussed these and related topics in the following papers: "Minds and machines," in *Dimensions of Mind*, Sidney Hook, ed.,(New York: New York University Press, 1960), pp. 148–79; "Brains and behavior," in *Analytical Philosophy*, second series, Ronald Butler, ed., (Oxford: Oxford University Press, 1965), pp. 1–20; and "The Mental Life of Some Machines," in *Intentionality, Minds, and Perception*, Hector-Neri Castañeda, ed., (Detroit: Wayne State University Press, 1967), pp. 177–200.
2. In this paper I wish to avoid the vexed question of the relation between *pains* and *pain states*. I only remark in passing that one common argument *against* identification of these two—viz., that a pain can be in one's arm but a state (of the organism) cannot be in one's arm—is easily seen to be fallacious.
3. There are some well-known remarks by Alonzo Church on this topic. Those remarks do not bear (as might at first be supposed) on the identification of concepts with synonymy-classes as such, but rather support the view that (in formal semantics) it is necessary to retain Frege's distinction between the normal and the 'oblique' use of expressions. That is, even if we say that the concept of temperature *is* the synonymy-class of the word 'temperature,' we must not thereby be led into the error of supposing that 'the concept of temperature' is synonymous with 'the synonymy-class of the word 'temperature"—for then 'the concept of temperature' and 'der Begriff der Temperatur' would not be synonymous, which they are. Rather, we must say that 'the concept of temperature' *refers* to the synonymy-class of the word 'temperature' (on this particular reconstruction); but that class is *identified* not as 'the synonymy class to which such-and-such a word belongs,' but in another way (e.g., as the synonymy-class whose members have such-and-such a characteristic use).
4. In the 'Mental life of some machines' a further, and somewhat independent, characteristic of the pain inputs is discussed in terms of Automata models—namely the spontaneity of the inclination to withdraw the injured part, etc. This raises the question, which is discussed in that paper, of giving a functional analysis of the notion of a spontaneous inclination. Of course, still further characteristics come readily to mind—for example, that feelings of pain are (or seem to be) *located* in the parts of the body.
5. Cf. the discussion of 'super-spartans' in "Brains and behavior."

The Causal Theory of the Mind

David M. Armstrong

Is Philosophy Just Conceptual Analysis?

What can philosophy contribute to solving the problem of the relation to mind to body? Twenty years ago, many English-speaking philosophers would have answered: 'Nothing beyond an analysis of the various mental *concepts*.' If we seek knowledge of things, they thought, it is to science that we must turn. Philosophy can only cast light upon our concepts of those things.

This retreat from things to concepts was not undertaken lightly. Ever since the seventeenth century, the great intellectual fact of our culture has been the incredible expansion of knowledge both in the natural and in the rational sciences (mathematics, logic). Everyday life presents us with certain simple verities. But, it seems, through science and only through science can we build upon these verities, and with astonishing results.

The success of science created a crisis in philosophy. What was there for philosophy to do? Hume had already perceived the problem in some degree, and so surely did Kant, but it was not until the twentieth century, with the Vienna Circle and with Wittgenstein, that the difficulty began to weigh heavily. Wittgenstein took the view that philosophy could do no more than strive to undo the intellectual knots it itself had tied, so achieving intellectual release, and even a certain illumination, but no knowledge. A little later, and more optimistically, Ryle saw a positive, if reduced, role for philosophy in mapping the 'logical geography' of our concepts: how they stood to each other and how they were to be analyzed.

On the whole, Ryle's view proved more popular than Wittgenstein's. After all, it retained a special, if much reduced, realm for philosophy where she might still be queen. There was better hope of continued employment for members of the profession!

Since that time, however, philosophers in the 'analytic' tradition have swung back from Wittgensteinian and even Rylean pessimism to a more traditional conception of the proper role and tasks of philosophy. Many analytic philosophers now would accept the view that the central task of philosophy is to give an account, or at least play a part in giving an account, of the most general nature of things and of man. (I would include myself among that many.)

Why has this swing back occurred? Has the old urge of the philosopher to determine the nature of things by a priori reasoning proved too strong? To use Freudian terms, are we simply witnessing a return of what philosophers had repressed? I think not. One consideration that has had great influence was the realization that those who thought that they were abandoning ontological and other substantive questions for a mere investigation of concepts were in fact smuggling in views on the substantive questions. They did not acknowledge that they held these views, but the views were there; and far worse from their standpoint, the views imposed a form upon their answers to the conceptual questions.

For instance, in *The Concept of Mind* (1949), Gilbert Ryle, although he denied that he was a Behaviorist, seemed to be upholding an account of man and his mind that was extremely close to Behaviorism. Furthermore, it seemed in many cases that it was this view of the mind–body problem that led him to his particular analyses of particular mental concepts, rather than the other way around. Faced with examples like this, it began to appear that, since philosophers could not help holding views on substantive matters, and the views could not help affecting their analyses of concepts, the views had better be held and discussed explicitly instead of appearing in a distorted, because unacknowledged, form.

The swing back by analytic philosophers to first-order questions was also due to the growth of a more sophisticated understanding of the nature of scientific investigation. For a philosophical tradition that is oriented towards science, as, on the whole, Western philosophy is, the consideration of the *methods* of science

From David M. Armstrong, *The Nature of Mind and Other Essays* (University of Queensland Press, 1981), pp. 16–31. Reprinted with permission of the publisher.

must be an important topic. It was gradually realized that in the past scientific investigation had regularly been conceived in far too positivistic, sensationalistic and observationalistic a spirit. (The influence of Karl Popper has been of the greatest importance in this realization.) As the central role of speculation, theory and reasoning in scientific investigation began to be appreciated by more and more philosophers, the border-line between science and philosophy began to seem at least more fluid, and the hope arose again that philosophy might have something to contribute to first-order questions.

The philosopher has certain special skills. These include the stating and assessing of the worth of arguments, including the bringing to light and making explicit suppressed premises of arguments, the detection of ambiguities and inconsistencies, and, perhaps especially, the analysis of concepts. But, I contend, these special skills do not entail that the *objective* of philosophy is to do these things. They are rather the special *means* by which philosophy attempts to achieve further objectives. Ryle was wrong in taking the analysis of concepts to be the end of philosophy. Rather, the analysis of concepts is a means by which the philosopher makes his contribution to great general questions, not about concepts, but about things.

In the particular case of the mind–body problem, the propositions the philosopher arrives at need not be of a special nature. They perhaps might have been arrived at by the psychologist, the neuro-physiologist, the biochemist or others, and, indeed, may be suggested to the philosopher by the results achieved or programs proposed by those disciplines. But the way that the argument is marshalled by a philosopher will be a special way. Whether this special way has or has not any particular value in the search for truth is a matter to be decided in particular cases. There is no a priori reason for thinking that the special methods of philosophy will be able to make a contribution to the mind–body problem. But neither is there an a priori reason for assuming that the philosopher's contribution will be valueless.

The Concept of a Mental State

The philosophy of philosophy is perhaps a somewhat joyless and unrewarding subject for reflection. Let us now turn to the mind–body problem itself, hoping that what is to be said about this particular topic will confirm the general remarks about philosophy that have just been made.

If we consider the mind-body problem today, then it seems that we ought to take account of the following consideration. The present state of scientific knowledge makes it probable that we can give a purely physico-chemical account of man's body. It seems increasingly likely that the body and the brain of man are constituted and work according to exactly the same principles as those physical principles that govern other, non-organic, matter. The differences between a stone and a human body appear to lie solely in the extremely complex material set-up that is to be found in the living body and which is absent in the stone. Furthermore, there is rather strong evidence that it is the state of our brain that completely determines the state of our consciousness and our mental state generally.

All this is not beyond the realm of controversy, and it is easy to imagine evidence that would upset the picture. In particular, I think that it is just possible that evidence from psychical research might be forthcoming that a physicochemical view of man's brain could not accommodate. But suppose that the physicochemical view of the working of the brain is correct, as I take it to be. It will be very natural to conclude that mental states are not simply *determined* by corresponding states of the brain, but that they are actually *identical* with these brain-states, brain-states that involve nothing but physical properties.

The argument just outlined is quite a simple one, and it hardly demands philosophical skill to develop it or to appreciate its force! But although many contemporary thinkers would accept its conclusion, there are others, including many philosophers, who would not. To a great many thinkers it has seemed obvious a priori that mental states could not be physical states of the brain. Nobody would identify a number with a piece of rock: it is sufficiently obvious that the two entities fall under different categories. In the same way, it has been thought, a perception or a feeling of sorrow must be a different category of thing from an electro-chemical discharge in the central nervous system.

Here, it seems to me, is a question to which philosophers can expect to make a useful contribution. It is a question about mental concepts. Is our concept of a mental state such that it is an intelligible hypothesis that mental states are physical states of the brain? If the philosopher

can show that it is an *intelligible* proposition (that is, a non-self-contradictory proposition) that mental states are physical states of the brain, then the scientific argument just given above can be taken at its face value as a strong reason for accepting the truth of the proposition.

My view is that the identification of mental states with physical states of the brain is a perfectly intelligible one, and that this becomes clear once we achieve a correct view of the analysis of the mental concepts. I admit that my analysis of the mental concepts was itself adopted because it permitted this identification, but such a procedure is commonplace in the construction of theories, and perfectly legitimate. In any case, whatever the motive for proposing the analysis, it is there to speak for itself, to be measured against competitors, and to be assessed as plausible or implausible independently of the identification it makes possible.

The problem of the identification may be put in a Kantian way: 'How is it possible that mental states should be physical states of the brain?' The solution will take the form of proposing an *independently plausible* analysis of the concept of a mental state that will permit this identification. In this way, the philosopher makes the way smooth for a first-order doctrine, which, true or false, is a doctrine of the first importance: a purely physicalist view of man.

The analysis proposed may be called the Causal analysis of the mental concepts. According to this view, the concept of a mental state essentially involves, and is exhausted by, the concept of a state that is *apt to be the cause of certain effects or apt to be the effect of certain causes.*

An example of a causal concept is the concept of poison. The concept of poison is the concept of something that when introduced into an organism causes that organism to sicken and/or die.[1] This is but a rough analysis of the concept the structure of which is in fact somewhat more complex and subtle than this. If *A* pours molten lead down *B*'s throat, then he may cause *B* to die as a result, but he can hardly be said to have poisoned him. For a thing to be called a poison, it is necessary that it act in a certain *sort* of way: roughly, in a biological as opposed to a purely physical way. Again, a poison can be introduced into the system of an organism and that organism fail to die or even to sicken. This might occur if an antidote were administered promptly. Yet again, the poison may be present in insufficient quantities to do any damage. Other qualifications could be made.

But the essential point about the concept of poison is that it is the concept of *that, whatever it is, which produces certain effects.* This leaves open the possibility of the *scientific identification* of poisons, of discovering that a certain sort of substance, such as cyanide, is a poison, and discovering further what it is about the substance that makes it poisonous.

Poisons are accounted poisons in virtue of their active powers, but many sorts of thing are accounted the sorts of thing they are by virtue of their *passive* powers. Thus brittle objects are accounted brittle because of the disposition they have to break and shatter when sharply struck. This leaves open the possibility of discovering empirically what sorts of thing are brittle and what it is about them that makes them brittle.

Now *if* the concepts of the various sorts of mental state are concepts of that which is, in various sorts of ways, apt for causing certain effects and apt for being the effect of certain causes, then it would be a quite unpuzzling thing if mental states should turn out to be physical states of the brain.

The concept of a mental state is the concept of something that is, characteristically, the cause of certain effects and the effect of certain causes. What sort of effects and what sort of causes? The effects caused by the mental state will be certain patterns of behavior of the person in that state. For instance, the desire for food is a state of a person or animal that characteristically brings about food-seeking and food-consuming behavior by that person or animal. The causes of mental states will be objects and events in the person's environment. For instance, a sensation of green is the characteristic effect in a person of the action upon his eyes of a nearby green surface.

The general pattern of analysis is at its most obvious and plausible in the case of *purposes.* If a man's purpose is to go to the kitchen to get something to eat, it is completely natural to conceive of this purpose as a cause within him that brings about, or tends to bring about, that particular line of conduct. It is, furthermore, notorious that we are unable to characterize purposes *except* in terms of that which they tend to bring about. How can we distinguish the purpose to go to the kitchen to get something to eat from another purpose to go to the bedroom to lie down? Only by the different outcomes that the two purposes tend to bring about. This fact was an encouragement to Behaviorism. It is still more plausibly explained by saying that the concept of purpose is a causal concept. The

further hypothesis that the two purposes are, in their own nature, different physical patterns in, or physical states of, the central nervous system is then a natural (although, of course, not logically inevitable) supplement to the causal analysis.

Simple models have great value in trying to grasp complex conceptions, but they are ladders that may need to be kicked away after we have mounted up by their means. It is vital to realize that the mental concepts have a far more complex logical structure than simple causal notions such as the concept of poison. The fact should occasion no surprise. In the case of poisons, the effect of which they are the cause is a gross and obvious phenomenon and the level of causal explanation involved in simply calling a substance 'a poison' is crude and simple. But in the case of mental states, their effects are all those complexities of behavior that mark off men and higher animals from the rest of the objects in the world. Furthermore, differences in such behavior are elaborately correlated with differences in the mental causes operating. So it is only to be expected that the causal patterns invoked by the mental concepts should be extremely complex and sophisticated.

In the case of the notion of a purpose, for instance, it is plausible to assert that it is the notion of a cause within which drives, or tends to drive, the man or animal through a series of actions to a certain end-state. But this is not the whole story. A purpose is only a purpose if it works to bring about behavioral effects *in a certain sort of way.* We may sum up this sort of way by saying that purposes are *information-sensitive* causes. By this is meant that purposes direct behavior by utilizing *perceptions* and *beliefs*, perceptions and beliefs about the agent's current situation and the way it develops, and beliefs about the way the world works. For instance, it is part of what it is to be a purpose to achieve *X* that this cause will cease to operate, will be 'switched off,' if the agent perceives or otherwise comes to believe that *X* has been achieved.

At this point, we observe that an account is being given of that special species of cause that is a purpose in terms of *further* mental items: perceptions and beliefs. This means that if we are to give a purely causal analysis even of the concept of a purpose we also will have to give a purely causal analysis of perceptions and beliefs. We may think of man's behavior as brought about by the joint operation of two sets of causes: first, his purposes and, second, his perceptions of and/or beliefs about the world. But since perceptions and beliefs are quite different sorts of thing from purposes, a Causal analysis must assign quite different causal *roles* to these different things in the bringing about of behavior.

I believe that this can be done by giving an account of perceptions and beliefs as *mappings* of the world. They are structures within us that model the world beyond the structure. This model is created in us by the world. Purposes may then be thought of as driving causes that utilize such mappings.

This is a mere thumb-nail, which requires much further development as well as qualification. One point that becomes clear when that development is given is that just as the concept of purpose cannot be elucidated without appealing to the concepts of perception and belief, so the latter cannot be elucidated without appealing to the concept of purpose. (This comes out, for instance, when we raise Hume's problem: what marks off beliefs from the mere entertaining of the same proposition? It seems that we can only mark off beliefs as those mappings in the light of which we are prepared to *act*, that is, which are potential servants of our purposes.) The logical dependence of purpose on perception and belief, and of perception and belief upon purpose, is not circularity in definition. What it shows is that the corresponding concepts *must be introduced together or not at all.* In itself, there is nothing very surprising in this. Correlative or mutually implicated concepts are common enough: for instance, the concepts of husband and wife or the concepts of soldier and army. No husbands without wives or wives without husbands. No soldiers without an army, no army without soldiers. But if the concepts of purpose, perception and belief are (i) correlative concepts and (ii) different species of purely causal concepts, then it is clear that they are far more complex in structure than a simple causal concept like poison. What falls under the mental concepts will be a complex and interlocking set of causal factors, which together are responsible for the 'minded' behavior of men and the higher animals.

The working out of the Causal theory of the mental concepts thus turns out to be an extremely complex business. Indeed when it is merely baldly stated, the Causal theory is, to use the phrase of Imre Lakatos, a *research program* in conceptual analysis rather than a developed theory. I have tried to show that it is a hopeful program by attempting, at least in

outline, a Causal analysis of all the main concepts in *A Materialist Theory of Mind* (1968); and I have supplemented the rather thin account given there of the concepts of belief, knowledge and inferring in *Belief, Truth and Knowledge* (1973).

Two examples of mental concepts where an especially complex and sophisticated type of Causal analysis is required are the notions of introspective awareness (one sense of the word 'consciousness') and the having of mental imagery. Introspective awareness is analyzable as a mental state that is a 'perception' of mental states. It is a mapping of the causal factors themselves. The having of mental imagery is a sort of mental state that cannot be elucidated in *directly* causal terms, but only by resemblance to the corresponding perceptions, which *are* explicated in terms of their causal role.

Two advantages of the Causal theory may now be mentioned. First, it has often been remarked by philosophers and others that the realm of mind is a shadowy one, and that the nature of mental states is singularly elusive and hard to grasp. This has given aid and comfort to Dualist or Cartesian theories of mind, according to which minds are quite different sorts of thing from material objects. But if the Causal analysis is correct, the facts admit of another explanation. What Dualist philosophers have grasped in a confused way is that our direct acquaintance with mind, which occurs in introspective awareness, is an acquaintance with something that we are aware of only as something that is causally linked, directly or indirectly, with behavior. In the case of our purposes and desires, for instance, we are often (though not invariably) introspectively aware of them. What we are aware of is the presence of factors within us that drive in a certain direction. We are not aware of the intrinsic nature of the factors. This emptiness or gap in our awareness is then interpreted by Dualists as immateriality. In fact, however, if the Causal analysis is correct, there is no warrant for this interpretation and, if the Physicalist identification of the nature of the causes is correct, the interpretation is actually false.

Second, the Causal analysis yields a still more spectacular verification. It shows promise of explaining a philosophically notorious feature of all or almost all mental states: their *intentionality*. This was the feature of mental states to which Brentano in particular drew attention, the fact that they may point towards certain objects or states of affairs, but that these objects and states of affairs need not exist. When a man strives, his striving has an objective, but that objective may never be achieved. When he believes, there is something he believes, but what he believes may not be the case. This capacity of mental states to 'point' to what does not exist can seem very special. Brentano held that intentionality set the mind completely apart from matter.

Suppose, however, that we consider a concept like the concept of poison. Does it not provide us with a miniature and unsophisticated model for the intentionality of mental states? Poisons are substances apt to make organisms sicken and die when the poison is administered. So it may be said that this is what poisons 'point' to. Nevertheless, poisons may fail of their effect. A poison does not fail to be a poison because an antidote neutralizes the customary effect of the poison.

May not the intentionality of mental states, therefore, be in principle a no more mysterious affair, although indefinitely more complex, than the death that lurks in the poison? As an intermediate case between poisons and mental states, consider the mechanisms involved in a homing rocket. Given a certain setting of its mechanism, the rocket may 'point' towards a certain target in a way that is a simulacrum of the way in which purposes point towards their objectives. The mechanism will only bring the rocket to the target in 'standard' circumstances: many factors can be conceived that would 'defeat' the mechanism. For the mechanism to operate successfully, some device will be required by which the developing situation is 'mapped' in the mechanism (i.e. what course the rocket is currently on, etc.). This mapping is an elementary analogue of perception, and so the course that is 'mapped' in the mechanism may be thought of as a simulacrum of the perceptual intentional object. Through one circumstance or another (e.g. malfunction of the gyroscope) this mapping may be 'incorrect.'

It is no objection to this analogy that homing rockets are built by men with purposes, who deliberately stamp a crude model of their own purposes into the rocket. Homing rockets might have been natural products, and non-minded objects that operate in a similar but far more complex way are found in nature. The living cell is a case in point.

So the Causal analyses of the mental concepts show promise of explaining both the transparency and the intentionality of mental states. One problem quite frequently raised in connection with these analyses, however, is in what sense they can be called 'analyses.' The

welter of complications in which the so-called analyses are involved make it sufficiently obvious that they do not consist of *synonymous translations* of statements in which mental terms figure. But, it has been objected, if synonymous translations of mental statements are unavailable, what precisely can be meant by speaking of 'analyses of concepts'?

I am far from clear what should be said in reply to this objection. Clearly, however, it does depend upon taking all conceptual analyses as claims about the synonymy of sentences, and that seems to be too simple a view. Going back to the case of poison: it is surely not an empirical fact, to be learnt by experience, that poisons kill. It is at the center of our notion of what poisons are that they have the power to bring about this effect. If they did not do that, they would not be properly called 'poisons.' But although this seems obvious enough, it is extremely difficult to give exact translations of sentences containing the word 'poison' into other sentences that do not contain the word or any synonym. Even in this simple case, it is not at all clear that the task can actually be accomplished.

For this reason, I think that sentence translation (with synonymy) is too strict a demand to make upon a purported conceptual analysis. What more relaxed demand can we make and still have a conceptual analysis? I do not know. One thing that we clearly need further light upon here is the concept of a concept, and how concepts are tied to language. I incline to the view that the connection between concepts and language is much less close than many philosophers have assumed. Concepts are linked primarily with belief and thought, and belief and thought, I think, have a great degree of logical independence of language, however close the empirical connection may be in many cases. If this is so, then an analysis of concepts, although of course conducted *in* words, may not be an investigation *into* words. (A compromise proposal: analysis of concepts might be an investigation into some sort of 'deep structure'—to use the currently hallowed phrase—which underlies the use of certain words and sentences.) I wish I were able to take the topic further.

The Problem of the Secondary Qualities

No discussion of the Causal theory of the mental concepts is complete that does not say something about the *secondary qualities*. If we consider such mental states as purposes and intentions, their 'transparency' is a rather conspicuous feature. It is notorious that introspection cannot differentiate such states except in terms of their different objects. It is not so immediately obvious, however, that *perception* has this transparent character. Perception involves the experience of color and of visual extension; touch the experience of the whole obscure range of tactual properties, including tactual extension; hearing, taste and smell the experience of sounds, tastes and smells. These phenomenal qualities, it may be argued, endow different perceptions with different qualities. The lack of transparency is even more obvious in the case of bodily sensations. Pains, itches, tickles and tingles are mental states, even if mental states of no very high-grade sort, and they each seem to involve their own peculiar qualities. Again, associated with different emotions it is quite plausible to claim to discern special emotion qualities. If perception, bodily sensation and emotions involve qualities, then this seems to falsify a purely Causal analysis of these mental states. They are not mere 'that whiches' known only by their causal role.

However, it is not at all clear how strong is the line of argument sketched in the previous paragraph. We distinguish between the intention and what is intended, and in just the same way we must distinguish between the perception and what is perceived. The intention is a mental state and so is the perception, but what is intended is not in general something mental and nor is what is perceived. What is intended may not come to pass, it is a merely intentional object, and the same may be said of what is perceived. Now in the case of the phenomenal qualities, it seems plausible to say that they are qualities not of the perception but rather of what is perceived. 'Visual extension' is the shape, size, etc. that some object of visual perception is perceived to have (an object that need not exist). Color seems to be a quality of that object. And similarly for the other phenomenal qualities. Even in the case of the bodily sensations, the qualities associated with the sensations do not *appear* to be qualities of mental states but instead to be qualities of portions of our bodies: more or less fleeting qualities that qualify the place where the sensation is located. Only in the case of the emotions does it seem natural to place the quality on the mental rather than the object side: but then it is not so clear whether there really *are* peculiar qualities associated with the emotions. The different patterns

of bodily sensations associated with the different emotions may be sufficient to do phenomenological justice to the emotions.

For these reasons, it is not certain whether the phenomenal qualities pose any threat to the Causal analysis of the mental concepts. But what a subset of these qualities quite certainly does pose a threat to, is the doctrine that the Causal analysis of the mental concepts is a step towards: Materialism or Physicalism.

The qualities of colour, sound, heat and cold, taste and smell together with the qualities that appear to be involved in bodily sensations and those that may be involved in the case of the emotions, are an embarrassment to the modern Materialist. He seeks to give an account of the world and of man purely in terms of *physical* properties, that is to say in terms of the properties that the physicist appeals to in his explanations of phenomena. The Materialist is not committed to the *current* set of properties to which the physicist appeals, but he is committed to whatever set of properties the physicist in the end will appeal to. It is clear that such properties as color, sound, taste and smell—the so-called 'secondary qualities'—will never be properties to which the physicist will appeal.

It is, however, a plausible thesis that associated with different secondary qualities are properties that are respectable from a physicist's point of view. Physical surfaces *appear* to have color. They not merely appear to, but undoubtedly do, emit light-waves, and the different mixtures of lengths of wave emitted are linked with differences in color. In the same way, different sorts of sound are linked with different sorts of sound-wave and differences in heat with differences in the mean kinetic energy of the molecules composing the hot things. The Materialist's problem therefore would be very simply solved if the secondary qualities could be identified with these physically respectable properties. (The qualities associated with bodily sensations would be identified with different sorts of stimulation of bodily receptors. If there are unique qualities associated with the emotions, they would presumably be identified with some of the physical states of the brain linked with particular emotions.)

But now the Materialist philosopher faces a problem. Previously he asked: 'How is it possible that mental states could be physical states of the brain?' This question was answered by the Causal theory of the mental concepts. Now he must ask: 'How is it possible that secondary qualities could be purely physical properties of the objects they are qualities of?' A Causal analysis does not seem to be of any avail. To try to give an analysis of, say, the quality of being red in Causal terms would lead us to produce such analyses as 'those properties of a physical surface, whatever they are, that characteristically produce *red sensations* in us.' But this analysis simply shifts the problem unhelpfully from property of surface to property of sensation. Either the red sensations involve nothing but physically respectable properties or they involve something more. If they involve something more, Materialism fails. But if they are simply physical states of the brain, having nothing but physical properties, then the Materialist faces the problem: 'How is it possible that red sensations should be physical states of the brain?' This question is no easier to answer than the original question about the redness of physical surfaces. (To give a Causal analysis of red sensations as the characteristic effects of the action of red surfaces is, of course, to move round in a circle.)

The great problem presented by the secondary qualities, such as redness, is that they are *unanalyzable.* They have certain relations of resemblance and so on to each other, so they cannot be said to be completely simple. But they are simple in the sense that they resist any analysis. You cannot give any complete account of the concept of redness without involving the notion of redness itself. This has seemed to be, and still seems to many philosophers to be, an absolute bar to identifying redness with, say, certain patterns of emission of light-waves.

But I am not so sure. I think it can be maintained that although the secondary qualities *appear* to be simple, they are not in fact simple. Perhaps their simplicity is *epistemological* only, not ontological, a matter of our awareness of them rather than the way they are. The best model I can give for the situation is the sort of phenomena made familiar to us by the Gestalt psychologists. It is possible to grasp that certain things or situations have a certain special property, but be unable to analyze that property. For instance, it may be possible to perceive that certain people are all alike in some way without being able to make it clear to oneself what the likeness is. We are aware that all these people have a certain likeness to each other, but are unable to define or specify that likeness. Later psychological research may achieve a specification of the likeness, a specification that may come as a complete surprise to us. Perhaps, therefore, the secondary qualities are in fact

complex, and perhaps they are complex characteristics of a sort demanded by Materialism, but we are unable to grasp their complexity in perception.

There are two divergences between the model just suggested and the case of the secondary qualities. First, in the case of grasping the indefinable likeness of people, we are under no temptation to think that the likeness is a likeness in some simple quality. The likeness is indefinable, but we are vaguely aware that it is complex. Second, once research has determined the concrete nature of the likeness, our attention can be drawn to, and we can observe individually, the features that determine the likeness.

But although the model suggested and the case of the secondary qualities undoubtedly exhibit these differences, I do not think that they show that the secondary qualities cannot be identified with respectable physical characteristics of objects. Why should not a complex property appear to be simple? There would seem to be no contradiction in adding such a condition to the model. It has the consequence that perception of the secondary qualities involves an element of illusion, but the consequence involves no contradiction. It is true also that in the case of the secondary qualities the illusion cannot be overcome within perception: it is impossible to see a colored surface as a surface emitting certain light-waves. (Though one sometimes seems to *hear* a sound as a vibration of the air.) But while this means that the identification of color and light-waves is a purely *theoretical* one, it still seems to be a possible one. And if the identification is a possible one, we have general scientific reasons to think it a *plausible* tone.

The doctrine of mental states and of the secondary qualities briefly presented in this paper seems to me to show promise of meeting many of the traditional philosophical objections to a Materialist or Physicalist account of the world. As I have emphasized, the philosopher is not professionally competent to argue the positive case for Materialism. There he must rely upon the evidence presented by the scientist, particularly the physicist. But at least he may neutralize the objections to Materialism advanced by his fellow philosophers.

NOTE

1. 'Any substance which, when introduced into or absorbed by a living organism, destroys life or injures health.' (*Shorter Oxford Dictionary*, 3rd ed., rev., 1978).

Mad Pain and Martian Pain

David Lewis

I

There might be a strange man who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its causes and effects. Our pain is typically caused by cuts, burns, pressure, and the like; his is caused by moderate exercise on an empty stomach. Our pain is generally distracting; his turns his mind to mathematics, facilitating concentration on that but distracting him from anything else. Intense pain has no tendency whatever to cause him to groan or writhe, but does cause him to cross his legs and snap his fingers. He is not in the least motivated to prevent pain or to get rid of it. In short, he feels pain but his pain does not at all occupy the typical causal role of pain. He would doubtless seem to us to be some sort of madman, and that

From Ned Block, ed., *Readings in the Philosophy of Psychology* (Harvard University Press, 1980).

is what I shall call him, though of course the sort of madness I have imagined may bear little resemblance to the real thing.

I said there might be such a madman. I don't know how to prove that something is possible, but my opinion that this is a possible case seems pretty firm. If I want a credible theory of mind, I need a theory that does not deny the possibility of mad pain. I needn't mind conceding that perhaps the madman is not in pain in *quite* the same sense that the rest of us are, but there had better be some straightforward sense in which he and we are both in pain.

Also, there might be a Martian who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its physical realization. His hydraulic mind contains nothing like our neurons. Rather, there are varying amounts of fluid in many-inflatable cavities, and the inflation of any one of these cavities opens some valves and closes others. His mental plumbing pervades most of his body—in fact, all but the heat exchanger inside his head. When you pinch his skin you cause no firing of C-fibers—he has none—but, rather, you cause the inflation of many smallish cavities in his feet. When these cavities are inflated, he is in pain. And the effects of his pain are fitting: his thought and activity are disrupted, he groans and writhes, he is strongly motivated to stop you from pinching him and to see to it that you never do again. In short, he feels pain but lacks the bodily states that either are pain or else accompany it in us.

There might be such a Martian; this opinion too seems pretty firm. A credible theory of mind had better not deny the possibility of Martian pain. I needn't mind conceding that perhaps the Martian is not in pain in *quite* the same sense that we Earthlings are, but there had better be some straightforward sense in which he and we are both in pain.

II

A credible theory of mind needs to make a place both for mad pain and for Martian pain. Prima facie, it seems hard for a materialist theory to pass this twofold test. As philosophers, we would like to characterize pain a priori. (We might settle for less, but let's start by asking for all we want.) As materialists, we want to characterize pain as a physical phenomenon. We can speak of the place of pain in the causal network from stimuli to inner states to behavior. And we can speak of the physical processes that go on when there is pain and that take their place in that causal network. We seem to have no other resources but these. But the lesson of mad pain is that pain is associated only contingently with its causal role, while the lesson of Martian pain is that pain is connected only contingently with its physical realization. How can we characterize pain a priori in terms of causal role and physical realization, and yet respect both kinds of contingency?

A simple identity theory straightforwardly solves the problem of mad pain. It goes just as straightforwardly wrong about Martian pain. A simple behaviorism or functionalism goes the other way: right about the Martian, wrong about the madman. The theories that fail our twofold test so decisively are altogether too simple. (Perhaps they are too simple ever to have had adherents.) It seems that a theory that can pass our test will have to be a mixed theory. It will have to be able to tell us that the madman and the Martian are both in pain, but for different reasons: the madman because he is in the right physical state, the Martian because he is in a state rightly situated in the causal network.

Certainly we can cook up a mixed theory. Here's an easy recipe. First, find a theory to take care of the common man and the madman, disregarding the Martian—presumably an identity theory. Second, find a theory to take care of the common man and the Martian, disregarding the madman—presumably some sort of behaviorism or functionalism. Then disjoin the two: say that to be in pain is to be in pain either according to the first theory or according to the second. Alternatively, claim ambiguity: say that to be in pain in one sense is to be in pain according to the first theory, to be in pain in another sense is to be in pain according to the second theory.

This strategy seems desperate. One wonders why we should have a disjunctive or ambiguous concept of pain, if common men who suffer pain are always in pain according to both disjuncts or both disambiguations. It detracts from the credibility of a theory that it posits a useless complexity in our concept of pain—useless in application to the common man, at least, and therefore useless almost always.

I don't object to the strategy of claiming ambiguity. As you'll see, I shall defend a version of it. But it's not plausible to cook up an ambiguity ad hoc to account for the compossibility of mad pain and Martian pain. It would be better to find a widespread sort of ambiguity, a sort we would believe in no matter what we

thought about pain, and show that it will solve our problem. That is my plan.

III

A dozen years or so ago, D. M. Armstrong and I (independently) proposed a materialist theory of mind that joins claims of type-type psychophysical identity with a behaviorist or functionalist way of characterizing mental states such as pain.[1] I believe our theory passes the twofold test. Positing no ambiguity without independent reason, it provides natural senses in which both madman and Martian are in pain. It wriggles through between Scylla and Charybdis.

Our view is that the concept of pain, or indeed of any other experience or mental state, is the concept of a state that occupies a certain causal role, a state with certain typical causes and effects. It is the concept of a state apt for being caused by certain stimuli and apt for causing certain behavior. Or, better, of a state apt for being caused in certain ways by stimuli plus other mental states and apt for combining with certain other mental states to jointly cause certain behavior. It is the concept of a member of a system of states that together more or less realize the pattern of causal generalizations set forth in common-sense psychology. (That system may be characterized as a whole and its members characterized afterward by reference to their place in it.)

If the concept of pain is the concept of a state that occupies a certain causal role, then whatever state does occupy that role is pain. If the state of having neurons hooked up in a certain way and firing in a certain pattern is the state properly apt for causing and being caused, as we materialists think, then that neural state is pain. But the concept of pain is not the concept of that neural state. ('The concept of . . .' is an intensional functor.) The concept of pain, unlike the concept of that neural state which in fact is pain, would have applied to some different state if the relevant causal relations had been different. Pain might have not been pain. The occupant of the role might have not occupied it. Some other state might have occupied it instead. Something that is not pain might have been pain.

This is not to say, of course, that it might have been that pain was not pain and nonpain was pain; that is, that it might have been that the occupant of the role did not occupy it and some nonoccupant did. Compare: 'The winner might have lost' (true) versus 'It might have been that the winner lost' (false). No wording is entirely unambiguous, but I trust my meaning is clear.

In short, the concept of pain as Armstrong and I understand it is a *nonrigid* concept. Likewise the word 'pain' is a nonrigid designator. It is a contingent matter what state the concept and the word apply to. It depends on what causes what. The same goes for the rest of our concepts and ordinary names of mental states.

Some need hear no more. The notion that mental concepts and names are nonrigid, wherefore what *is* pain might not have been, seems to them just self-evidently false.[2] I cannot tell why they think so. Bracketing my own theoretical commitments, I think I would have no opinion one way or the other. It's not that I don't care about shaping theory to respect naive opinion as well as can be, but in this case I have no naive opinion to respect. If I am not speaking to your condition, so be it.

If pain is identical to a certain neural state, the identity is contingent. Whether it holds is one of the things that varies from one possible world to another. But take care. I do not say that here we have two states, pain and some neural state, that are contingently identical, identical at this world but different at another. Since I'm serious about the identity, we have not two states but one. This one state, this neural state which is pain, is not contingently identical to itself. It does not differ from itself at any world. Nothing does.[3] What's true is, rather, that the concept and name of pain contingently apply to some neural state at this world, but do not apply to it at another. Similarly, it is a contingent truth that Bruce is our cat, but it's wrong to say that Bruce and our cat are contingently identical. Our cat Bruce is necessarily self-identical. What is contingent is that the nonrigid concept of being our cat applies to Bruce rather than to some other cat, or none.

IV

Nonrigidity might begin at home. All actualities are possibilities, so the variety of possibilities includes the variety of actualities. Though some possibilities are thoroughly otherworldly, others may be found on planets within range of our telescopes. One such planet is Mars.

If a nonrigid concept or name applies to different states in different possible cases, it should be no surprise if it also applies to different states in different actual cases. Nonrigidity

is to logical space as other relativities are to ordinary space. If the word 'pain' designates one state at our actual world and another at a possible world where our counterparts have a different internal structure, then also it may designate one state on Earth and another on Mars. Or, better, since Martians may come here and we may go to Mars, it may designate one state for Earthlings and another for Martians.

We may say that some state *occupies a causal role for a population.* We may say this whether the population is situated entirely at our actual world, or partly at our actual world and partly at other worlds, or entirely at other worlds. If the concept of pain is the concept of a state that occupies that role, then we may say that a state *is pain for a population.* Then we may say that a certain pattern of firing of neurons is pain for the population of actual Earthlings and some but not all of our otherworldly counterparts, whereas the inflation of certain cavities in the feet is pain for the population of actual Martians and some of their otherworldly counterparts. Human pain is the state that occupies the role of pain for humans. Martian pain is the state that occupies the same role for Martians.

A state occupies a causal role for a population, and the concept of occupant of that role applies to it, if and only if, with few exceptions, whenever a member of that population is in that state, his being in that state has the sort of causes and effects given by the role.

The thing to say about Martian pain is that the Martian is in pain because he is in a state that occupies the causal role of pain for Martians, whereas we are in pain because we are in a state that occupies the role of pain for us.

V

Now, what of the madman? He is in pain, but he is not in a state that occupies the causal role of pain for him. He is in a state that occupies that role for most of us, but he is an exception. The causal role of a pattern of firing of neurons depends on one's circuit diagram, and he is hooked up wrong.

His state does not occupy the role of pain for a population comprising himself and his fellow madmen. But it does occupy that role for a more salient population—mankind at large. He is a man, albeit an exceptional one, and a member of that larger population.

We have allowed for exceptions. I spoke of the definitive syndrome of *typical* causes and effects. Armstrong spoke of a state *apt for* having certain causes and effects; that does not mean that it has them invariably. Again, I spoke of a system of states that *comes near to* realizing common-sense psychology. A state may therefore occupy a role for mankind even if it does not at all occupy that role for some mad minority of mankind.

The thing to say about mad pain is that the madman is in pain because he is in the state that occupies the causal role of pain for the population comprising all mankind. He is an exceptional member of that population. The state that occupies the role for the population does not occupy it for him.

VI

We may say that *X* is in pain *simpliciter* if and only if *X* is in the state that occupies the causal role of pain for the *appropriate* population. But what is the appropriate population? Perhaps (1) it should be *us;* after all, it's our concept and our word. On the other hand, if it's *X* we're talking about, perhaps (2) it should be a population that *X* himself belongs to, and (3) it should preferably be one in which *X* is not exceptional. Either way, (4) an appropriate population should be a natural kind—a species, perhaps.

If *X* is you or I—human and unexceptional—all four considerations pull together. The appropriate population consists of mankind as it actually is, extending into other worlds only to an extent that does not make the actual majority exceptional.

Since the four criteria agree in the case of the common man, which is the case we usually have in mind, there is no reason why we should have made up our minds about their relative importance in cases of conflict. It should be no surprise if ambiguity and uncertainty arise in such cases. Still, some cases do seem reasonably clear.

If *X* is our Martian, we are inclined to say that he is in pain when the cavities in his feet are inflated; and so says the theory, provided that criterion (1) is outweighed by the other three, so that the appropriate population is taken to be the species of Martians to which *X* belongs.

If *X* is our madman, we are inclined to say that he is in pain when he is in the state that occupies the role of pain for the rest of us; and so says the theory, provided that criterion (3) is outweighed by the other three, so that the appropriate population is taken to be mankind.

We might also consider the case of a mad Martian, related to other Martians as the madman is to the rest of us. If X is a mad Martian, I would be inclined to say that he is in pain when the cavities in his feet are inflated; and so says our theory, provided that criteria (2) and (4) together outweigh either (1) or (3) by itself.

Other cases are less clear-cut. Since the balance is less definitely in favor of one population or another, we may perceive the relativity to population by feeling genuinely undecided. Suppose the state that plays the role of pain for us plays instead the role of thirst for a certain small subpopulation of mankind, and vice versa. When one of them has the state that is pain for us and thirst for him, there may be genuine and irresolvable indecision about whether to call him pained or thirsty—that is, whether to think of him as a madman or as a Martian. Criterion (1) suggests calling his state pain and regarding him as an exception; criteria (2) and (3) suggest shifting to a subpopulation and calling his state thirst. Criterion (4) could go either way, since mankind and the exceptional subpopulation may both be natural kinds. (Perhaps it is relevant to ask whether membership in the subpopulation is hereditary.)

The interchange of pain and thirst parallels the traditional problem of inverted spectra. I have suggested that there is no determinate fact of the matter about whether the victim of interchange undergoes pain or thirst. I think this conclusion accords well with the fact that there seems to be no persuasive solution one way or the other to the old problem of inverted spectra. I would say that there is a good sense in which the alleged victim of inverted spectra sees red when he looks at grass: he is in a state that occupies the role of seeing red for mankind in general. And there is an equally good sense in which he sees green: he is in a state that occupies the role of seeing green for him, and for a small subpopulation of which he is an unexceptional member and which has some claim to be regarded as a natural kind. You are right to say either, though not in the same breath. Need more be said?

To sum up. Armstrong and I claim to give a schema that, if filled in, would characterize pain and other states a priori. If the causal facts are right, then also we characterize pain as a physical phenomenon. By allowing for exceptional members of a population, we associate pain only contingently with its causal role. Therefore we do not deny the possibility of mad pain, provided there is not too much of it. By allowing for variation from one population to another (actual or merely possible) we associate pain only contingently with its physical realization. Therefore we do not deny the possibility of Martian pain. If different ways of filling in the relativity to population may be said to yield different senses of the word 'pain,' then we plead ambiguity. The madman is in pain in one sense, or relative to one population. The Martian is in pain in another sense, or relative to another population. (So is the mad Martian.)

But we do not posit ambiguity ad hoc. The requisite flexibility is explained simply by supposing that we have not bothered to make up our minds about semantic niceties that would make no difference to any commonplace case. The ambiguity that arises in cases of inverted spectra and the like is simply one instance of a commonplace kind of ambiguity—a kind that may arise whenever we have tacit relativity and criteria of selection that sometimes fail to choose a definite *relatum*. It is the same kind of ambiguity that arises if someone speaks of relevant studies without making clear whether he means relevance to current affairs, to spiritual well-being, to understanding, or what.

VII

We have a place for commonplace pain, mad pain, Martian pain, and even mad Martian pain. But one case remains problematic. What about pain in a being who is mad, alien, and unique? Have we made a place for that? It seems not. Since he is mad, we may suppose that his alleged state of pain does not occupy the proper causal role for him. Since he is alien, we may also suppose that it does not occupy the proper role for us. And since he is unique, it does not occupy the proper role for others of his species. What is left?

(One thing that might be left is the population consisting of him and his un-actualized counterparts at other worlds. If he went mad as a result of some improbable accident, perhaps we can say that he is in pain because he is in the state that occupies the role for most of his alternative possible selves, the state that would have occupied the role for him if he had developed in a more probable way. To make the problem as hard as possible, I must suppose that this solution is unavailable. He did *not* narrowly escape being so constituted that his present state would have occupied the role of pain.)

I think we cannot and need not solve this problem. Our only recourse is to deny that the case is possible. To stipulate that the being in this example is in pain was illegitimate. That seems credible enough. Admittedly, I might have thought offhand that the case was possible. No wonder; it merely combines elements of other cases that are possible. But I am willing to change my mind. Unlike my opinions about the possibility of mad pain and Martian pain, my naive opinions about this case are not firm enough to carry much weight.

VIII

Finally, I would like to try to preempt an objection. I can hear it said that I have been strangely silent about the very center of my topic. *What is it like* to be the madman, the Martian, the mad Martian, the victim of interchange of pain and thirst, or the being who is mad, alien, and unique? What is the *phenomenal character* of his state? If it *feels* to him like pain, then it *is* pain, whatever its causal role or physical nature. If not, it isn't. It's that simple!

Yes. It would indeed be a mistake to consider whether a state is pain while ignoring what it is like to have it. Fortunately, I have not made that mistake. Indeed, it is an impossible mistake to make. It is like the impossible mistake of considering whether a number is composite while ignoring the question of what factors it has.

Pain is a feeling.[4] Surely that is uncontroversial. To have pain and to feel pain are one and the same. For a state to be pain and for it to feel painful are likewise one and the same. A theory of what it is for a state to be pain is inescapably a theory of what it is like to be in that state, of how that state feels, of the phenomenal character of that state. Far from ignoring questions of how states feel in the odd cases we have been considering, I have been discussing nothing else! Only if you believe on independent grounds that considerations of causal role and physical realization have no bearing on whether a state is pain should you say that they have no bearing on how that state feels.

NOTES

This paper was presented at a conference on mind-body identity held at Rice University in April 1978. I am grateful to many friends, and especially to Patricia Kitcher, for valuable discussions of the topic.

1. D. M. Armstrong. *A Materialist Theory of the Mind* (London: Routledge, 1968); "The Nature of Mind," in *The Mind/Brain Identity Theory*, C. V. Borst, ed., (London: Macmillan, 1970), pp. 67–97; "The Causal Theory of the Mind," *Neue Heft für Philosophie*, no. 11 (Vendenhoek & Ruprecht, 1977), pp. 82–95. David Lewis, "An Argument for the Identity Theory," *Journal of Philosophy* 63 (1966): pp. 17–25, reprinted with additions in *Materialism and the Mind-Body Problem*, David M. Rosenthal, ed., (Englewood Cliffs, NJ: Prentice-Hall, 1971), pp. 162–71; "Review of *Art, Mind, and Religion*," *Journal of Philosophy* 66 (1969): pp. 22–27, particularly pp. 23–25; "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy* 50 (1972): pp. 249–58; "Radical Interpretation," *Synthese* 23 (1974): pp. 331–44.
2. For instance, see Saul A. Kripke, "Naming and Necessity," in *Semantics of Natural Language*, Gilbert Harman and Donald Davidson, eds., (Dordrecht: Reidel, 1972), pp. 253–355, 763–69, particularly pp. 335–36. Note that the sort of identity theory that Kripke opposes by argument, rather than by appeal to self-evidence, is not the sort that Armstrong and I propose.
3. The closest we can come is to have something at one world with twin counterparts at another. See my "Counterpart Theory and Quantified Modal Logic," *Journal of Philosophy* 65 (1968): pp. 113–26. That possibility is irrelevant to the present case.
4. Occurrent pain, that is. Maybe a disposition that sometimes, but not always, causes occurrent pain might also be called 'pain.'

Troubles with Functionalism

Ned Block

One characterization of functionalism that is probably vague enough to be accepted by most functionalists is: each type of mental state is a state consisting of a disposition to act in certain ways *and to have certain mental states*, given certain sensory inputs and certain mental states. So put, functionalism can be seen as a new incarnation of behaviorism. Behaviorism identifies mental states with dispositions to act in certain ways in certain input situations. But as critics have pointed out (Chisholm 1957; Putnam 1963), desire for goal G cannot be identified with, say, the disposition to do A in input circumstances in which A leads to G, since, after all, the agent might not *know* A leads to G and thus might not be disposed to do A. Functionalism replaces behaviorism's 'sensory inputs' with 'sensory inputs and mental states'; and functionalism replaces behaviorism's 'disposition to act' with 'disposition to act and have certain mental states.' Functionalists want to individuate mental states causally, and since mental states have mental causes and effects as well as sensory causes and behavioral effects, functionalists individuate mental states partly in terms of causal relations to other mental states. One consequence of this difference between functionalism and behaviorism is that there are organisms that according to behaviorism, have mental states but, according to functionalism, do not have mental states.

So, necessary conditions for mentality that are postulated by functionalism are in one respect stronger than those postulated by behaviorism. According to behaviorism, it is necessary and sufficient for desiring that G that a system be characterized by a certain set (perhaps infinite) of input-output relations; that is, according to behaviorism, a system desires that G just in case a certain set of conditionals of the form 'It will emit O given I' are true of it. According to functionalism, however, a system might have these input-output relations, yet not desire that G; for according to functionalism, whether a system desires that G depends on whether it has internal states which have certain causal relations to other internal states (and to inputs and outputs). Since behaviorism makes no such 'internal state' requirement, there are possible systems of which behaviorism affirms and functionalism denies that they have mental states.[1] One way of stating this is that, according to functionalism, behaviorism is guilty of *liberalism*—ascribing mental properties to things that do not in fact have them.

By 'physicalism,' I mean the doctrine that pain, for example, is identical to a physical (or physiological) state.[2] As many philosophers have argued (notably Fodor 1965, and Putnam 1966; see also Block and Fodor 1972), if functionalism is true, physicalism is false. The point is at its clearest with regard to Turing-machine versions of functionalism. Any given abstract Turing machine can be realized by a wide variety of physical devices; indeed, it is plausible that, given any putative correspondence between a Turing-machine state and a configurational physical (or physiological) state, there will be a possible realization of the Turing machine that will provide a counterexample to that correspondence. (See Kalke1969; Gendron 1971; Mucciolo1974, for unconvincing arguments to the contrary; see also Kim 1972.) Therefore, if pain is a functional state, it cannot, for example, be a brain state, because creatures without brains can realize the same Turing machine as creatures with brains.

One way of expressing this point is that, according to functionalism, physicalism is a *chauvinist* theory: it withholds mental properties from systems that in fact have them. In saying mental states are brain states, for example, physicalists unfairly exclude those poor brainless creatures who nonetheless have minds.

This chapter has three parts. The first [excerpted here—ed.] argues that functionalism is guilty of liberalism, the second that one way of modifying functionalism to avoid liberalism is to tie it more closely to empirical psychology,

Excerpted from C. W. Savage, ed., *Perception and Cognition* (University of Minnesota Press, 1978), pp. 261–325, with permission of the publisher.

and the third that no version of functionalism can avoid both liberalism and chauvinism.

1.1. More about What Functionalism Is

One can also categorize functionalists in terms of whether they regard functional identities as part of a priori psychology or empirical psychology. (Since this distinction crosscuts the machine/nonmachine distinction, I shall be able to illustrate nonmachine versions of functionalism in what follows.) The a priori functionalists (e.g., Smart, Armstrong, Lewis, Shoemaker) are the heirs of the logical behaviorists. They tend to regard functional analyses as analyses of the meanings of mental terms, whereas the empirical functionalists (e.g., Fodor, Putnam, Harman) regard functional analyses as substantive scientific hypotheses. In what follows, I shall refer to the former view as 'Functionalism' and the latter as 'Psychofunctionalism.' (I shall use 'functionalism' with a lowercase 'f' as neutral between Functionalism and Psychofunctionalism. When distinguishing between Functionalism and Psychofunctionalism, I shall always use capitals.)

Functionalism and Psychofunctionalism and the difference between them can be made clearer in terms of the notion of the Ramsey sentence of a psychological theory. Mental-state terms that appear in a psychological theory can be defined in various ways by means of the Ramsey sentence of the theory. All functional-state identity theories (and functional-property identity theories) can be understood as defining a set of functional states (or functional properties) by means of the Ramsey sentence of a psychological theory—with one functional state corresponding to each mental state (or one functional property corresponding to each mental property). The functional state corresponding to pain will be called the 'Ramsey functional correlate' of pain, with respect to the psychological theory. In terms of the notion of a Ramsey functional correlate with respect to a theory, the distinction between Functionalism and Psychofunctionalism can be defined as follows: Functionalism identifies mental state S with S's Ramsey functional correlate with respect to a *common-sense* psychological theory; Psychofunctionalism identifies S with S's Ramsey functional correlate with respect to a *scientific* psychological theory.

1.2. Homunculi-Headed Robots

In this section I shall describe a class of devices that embarrass all versions of functionalism in that they indicate functionalism is guilty of liberalism—classifying systems that lack mentality as having mentality.

Consider the simple version of machine functionalism already described. It says that each system having mental states is described by at least one Turing-machine table of a certain kind, and each mental state of the system is identical to one of the machine-table states specified by the machine table. I shall consider inputs and outputs to be specified by descriptions of neural impulses in sense organs and motor-output neurons. This assumption should not be regarded as restricting what will be said to Psychofunctionalism rather than Functionalism. As already mentioned, every version of functionalism assumes *some* specification of inputs and outputs. A Functionalist specification would do as well for the purposes of what follows.

Imagine a body externally like a human body, say yours, but internally quite different. The neurons from sensory organs are connected to a bank of lights in a hollow cavity in the head. A set of buttons connects to the motor-output neurons. Inside the cavity resides a group of little men. Each has a very simple task: to implement a 'square' of a reasonably adequate machine table that describes you. On one wall is a bulletin board on which is posted a state card, i.e., a card that bears a symbol designating one of the states specified in the machine table. Here is what the little men do: Suppose the posted card has a 'G' on it. This alerts the little men who implement G squares—'G-men' they call themselves. Suppose the light representing input I_{17} goes on. One of the G-men has the following as his sole task: when the card reads 'G' and the I_{17} light goes on, he presses output button O_{191} and changes the state card to 'M' This G-man is called upon to exercise his task only rarely. In spite of the low level of intelligence required of each little man, the system as a whole manages to simulate you because the functional organization they have been trained to realize is yours. A Turing machine can be represented as a finite set of quadruples (or quintuples, if the output is divided into two parts)—current state, current input; next state, next output. Each little man has the task corresponding to a single quadruple. Through the efforts of the little men, the

system realizes the same (reasonably adequate) machine table as you do and is thus functionally equivalent to you.

I shall describe a version of the homunculi-headed simulation, which is more clearly nomologically possible. How many homunculi are required? Perhaps a billion are enough; after all, there are only about a billion neurons in the brain.

Suppose we convert the government of China to functionalism, and we convince its officials that it would enormously enhance their international prestige to realize a human mind for an hour. We provide each of the billion people in China (I chose China because it has a billion inhabitants) with a specially designed two-way radio that connects them in the appropriate way to other persons and to the artificial body mentioned in the previous example. We replace the little men with a radio transmitter and receiver connected to the input and output neurons. Instead of a bulletin board, we arrange to have letters displayed on a series of satellites placed so that they can be seen from anywhere in China. Surely such a system is not physically impossible. It could be functionally equivalent to you for a short time, say an hour.

'But,' you may object, 'how could something be functionally equivalent to me for *an hour?* Doesn't my functional organization determine, say, how I would react to doing nothing for a week but reading *Reader's Digest?*' Remember that a machine table specifies a set of conditionals of the form: if the machine is in S_i and receives input I_j, it emits output O_k and goes into S_l. Any system that has a set of inputs, outputs, and states related in the way described realizes that machine table, even if it exists for only an instant. For the hour the Chinese system is 'on,' it *does* have a set of inputs, outputs, and states of which such conditionals are true. Whatever the initial state, the system will respond in whatever way the machine table directs. This is how *any* computer realizes the machine table it realizes.

Of course, there are signals the system would respond to that you would not respond to, e.g., massive radio interference or a flood of the Yangtze River. Such events might cause a malfunction, scotching the simulation, just as a bomb in a computer can make it fail to realize the machine table it was built to realize. But just as the computer *without* the bomb *can* realize the machine table, the system consisting of the people and artificial body can realize the machine table so long as there are no catastrophic interferences, e.g., floods, etc.

'But,' someone may object, 'there is a difference between a bomb in a computer and a bomb in the Chinese system, for in the case of the latter (unlike the former), inputs as specified in the machine table can be the cause of the malfunction. Unusual neural activity in the sense organs of residents of Chungking Province caused by a bomb or by a flood of the Yangtze can cause the system to go haywire.'

Reply: the person who says what system he or she is talking about gets to say what counts as inputs and outputs. I count as inputs and outputs only neural activity in the artificial body connected by radio to the people of China. Neural signals in the people of Chungking count no more as inputs to this system than input tape jammed by a saboteur between the relay contacts in the innards of a computer count as an input to the computer.

Of course, the object consisting of the people of China + the artificial body has *other* Turing machine descriptions under which neural signals in the inhabitants of Chungking *would* count as inputs. Such a new system (i.e., the object under such a new Turing-machine description) would not be functionally equivalent to you. Likewise, any commercial computer can be redescribed in a way that allows tape jammed into its innards to count as inputs. In describing an object as a Turing machine, one draws a line between the inside and the outside. (If we count only neural impulses as inputs and outputs, we draw that line inside the body if *we* count only peripheral stimulations as inputs and only bodily movements as outputs, we draw that line at the skin.) In describing the Chinese system as a Turing machine, I have drawn the line in such a way that it satisfies a certain type of functional description—one that you *also* satisfy, and one that, according to functionalism, justifies attributions of mentality. Functionalism does not claim that every mental system has a machine table of a sort that justifies attributions of mentality with respect to *every* specification of inputs and outputs, but rather, only with respect to *some* specification.

Objection: The Chinese system would work too slowly. The kind of events and processes with which we normally have contact would pass by far too quickly for the system to detect them. Thus, we would be unable to converse with it, play bridge with it, etc.[3]

Reply: It is hard to see why the system's time scale should matter. What reason is there to believe that *your* mental operations could not be very much slowed down, yet remain mental

operations? Is it really contradictory or nonsensical to suppose we could meet a race of intelligent beings with whom we could communicate only by devices such as time-lapse photography? When we observe these creatures, they seem almost inanimate. But when we view the time-lapse movies, we see them conversing with one another. Indeed, we find they are saying that the only way they can make any sense of us is by viewing movies greatly slowed down. To take time scale as all important seems crudely behavioristic. Further, even if the timescale objection is right, I can elude it by retreating to the point that a homunculus-head that works in normal time is *metaphysically* possible, even if not nomologically possible. Metaphysical possibility is all my argument requires (see Section 1.3).[4]

What makes the homunculi-headed system (count the two systems as variants of a single system) just described a prima facie counter example to (machine) functionalism is that there is prima facie doubt whether it has any mental states at all—especially whether it has what philosophers have variously called 'qualitative states,' 'raw feels,' or 'immediate phenomenological qualities.' (You ask: What is it that philosophers have called qualitative states? I answer, only half in jest. As Louis Armstrong said when asked what jazz is, 'If you got to ask, you ain't never gonna get to know.') In Nagel's terms (1974), there is a prima facie doubt whether there is anything which it is like to be the homunculi-headed system.

The force of the prima facie counterexample can be made clearer as follows: Machine functionalism says that each mental state is identical to a machine-table state. For example, a particular qualitative state, Q, is identical to a machine-table state, S_q. But if there is nothing it is like to be the homunculi-headed system, it cannot be in Q even when it is in S_q. Thus, if there is prima facie doubt about the homunculi-headed system's mentality, there is prima facie doubt that $Q = S_q$, i.e., doubt that the kind of functionalism under consideration is true.[5] Call this argument the Absent Qualia Argument.

REFERENCES

Block, N. "Are absent qualia impossible?" *Philosophical Review* 89 (1980): pp. 257–74.

Block, N., and Fodor, J., "What psychological states are not," *Philosophical Review* 81 (1972): pp. 159–81.

Chisholm, Roderick. *Perceiving* (Ithaca, NY: Cornell University Press, 1957).

Fodor, J., "Explanations in psychology," in *Philosophy in America*, M. Black, ed., (London: Routledge & Kegan Paul, 1965).

Gendron, B., "On the relation of neurological and psychological theories: A critique of the hardware thesis," in *Boston studies in the philosophy of Science VIII*, R. C. Buck and R. S. Cohen, eds., (Dordrecht: Reidel, 1971).

Hempel, C., "Reduction: Ontological and linguistic facets," in *Essays in honor of Ernest Nagel,* S. Morgenbesser, P. Suppes, and M. White, eds., (New York: St. Martin's Press, 1970.)

Kalke, W. "What is wrong with Fodor and Putnam's functionalism?" *Nôus* 3 (1969): pp. 83–93.

Kim, J., "Phenomenal properties, psychophysical laws, and the identity theory," *The Monist* 56 (2) (1972): pp. 177–92.

Mucciolo, L. F. "The identity thesis and neuropsychology," *Nôus* 8 (1974): pp.327–42.

Nagel, T. "What is it like to be a bat?" *Philosophical Review* 83 (1974): pp. 435–50.

Putnam, H. "Brains and behavior" (1963). Reprinted as are all Putnam's articles referred to here (except "On properties") in *Mind, language and reality; philosophical papers*, vol. 2 (London: Cambridge University Press, 1975).

_____. "The mental life of some machines" (1966).

_____. "On properties," in *Mathematics, matter and method; philosophical papers*, vol. 1 (London: Cambridge University Press, 1970.)

Shoemaker, S. "Functionalism and qualia," *Philosophical studies* 27 (1975): pp. 271–315.

NOTES

1. The converse is also true.
2. State type, not state token. Throughout the chapter, I shall mean by 'physicalism' the doctrine that says each distinct type of mental state is identical to a distinct type of physical state, for example, pain (the universal) is a physical state. Token physicalism, on the other hand, is the (weaker) doctrine that each particular datable pain is a state of some physical type or other. Functionalism shows that type physicalism is false, but it does not show that token physicalism is false.

   By 'physicalism,' I mean *first order* physicalism, the doctrine that, e.g., the property of being in pain is a first-order (in the Russell-Whitehead sense) physical property. (A first-order property is one whose definition does not require quantification over properties; a second-order property is one whose definition requires quantification over first-order properties.) The claim that being in pain is a second-order physical property is actually a (physicalist) form of functionalism. See Putnam 1970.

'Physical property' could be defined for the purposes of this chapter as a property expressed by a predicate of some true physical theory or, more broadly, by a predicate of some true theory of physiology, biology, chemistry, or physics. Of course, such a definition is unsatisfactory without characterizations of these branches of science. See Hempel 1970 for further discussion of this problem.

3. This point has been raised with me by persons too numerous to mention.
4. One potential difficulty for Functionalism is provided by the possibility that one person may have two radically different Functional descriptions of the sort that justify attribution of mentality. In such a case, Functionalists might have to ascribe two radically different systems of belief, desire, etc., to the same person, or suppose that there is no fact of the matter about what the person's propositional attitudes are. Undoubtedly, Functionalists differ greatly on what they make of this possibility, and the differences reflect positions on such issues as indeterminacy of translation.
5. Shoemaker 1975 argues (in reply to Block and Fodor 1972) that absent qualia are logically impossible, that is, that it is logically impossible that two systems be in the same functional state yet one's state have and the other's state lack qualitative content. If Shoemaker is right, it is wrong to doubt whether the homunculi-headed system has qualia. I attempt to show Shoemaker's argument to be fallacious in Block 1980.

Pseudonormal Vision

An Actual Case of Qualia Inversion?

Martine Nida-Rümelin

1. Introduction

Is it possible that a person who behaves just like you and me in normal life situations and applies colour words to objects just as we do and makes the same colour discriminations and colour similarity judgements that we make, see green where we see red and red where we see green? Many philosophers assert that the description of such a case is somehow incoherent. Often the motivation for this assertion is 'that they suspect that admitting that claim [the possibility of such a case] will put one on a slippery slope which will eventually land one in skepticism about other minds.'[1]

Among philosophers, however, it does not seem to be common knowledge that there is scientific evidence for the existence of such cases. Theories about the physiological basis of colour vision deficiencies together with theories about the genetics of colour vision deficiencies lead to the prediction that some people are 'pseudonormal' (according to an estimation of Piantanida (1974) this occurs in around 14 of 10,000 males).[2] Pseudonormal people 'would be expected to have normal colour vision except that the sensations of red and green would be reversed—something that would be difficult, if not impossible, to prove.'[3]

Any philosophical theory of mind or more specifically about colour, colour appearances or colour concepts should meet the following plausible prima facie constraint: *No hypotheses accepted or seriously considered in colour vision science should be regarded according to a philosophical theory to be either incoherent or unstatable or false.* Therefore—regardless of whether the hypothesis of the existence of pseudonormal people is correct—the mere fact that the hypothesis is seriously considered in colour vision science, is philosophically relevant. Central claims of colour vision science when combined with specific empirical assumptions lead to the prediction that there are red-green-inverted people. Therefore any philosophical theory which excludes such a case does not meet the above formulated constraint. The failure to meet this prima facie constraint does not in itself justify the rejection of a philosophical proposal, but it does represent a serious objection. This kind of criticism will be advanced against some widely held philosophical proposals in the present paper. But let me begin with a short sketch of the relevant parts of colour vision science.

From *Philosophical Studies* 82 (1996): pp. 145–57. Reprinted with permission of the author and of Kluwer Academic Publishers.

2. Pseudonormal Vision. The Scientific Background

There are three types of photoreceptors on the retina that play a central role in human colour vision (B-, G- and R-cones). They are morphologically distinguishable, they play different roles in colour information processing and they normally contain three chemically different photopigments. For each cone type there is a characteristic function (the so-called sensitivity curve) which describes how the level of stimulation caused by monochromatic light in a cone of the given type depends on the wavelengths of the light at a given intensity level. It is assumed that the sensitivity curves are determined by the absorption spectra of the pigments contained in the receptors. The expected level of stimulation of a cone caused by non-monochromatic light (which is the normal case) can be calculated on the basis of the sensitivity curve characteristic for its type. When light reaches a given area on the retina, then some neural mechanism will calculate the average stimulation of the cones in the area of any of the three types. The average stimulation of the three cone types is then compared and information about the results is carried to the brain by two neural channels, the r-g-channel (responsible for red- and green-sensations) and the y-b-channel (responsible for blue- and yellow-sensations). If (b, g, r) represents the average stimulation of the B-, G- and R-cones in the area at issue then how the channel states depend on the average stimulation (b, g, r) of the three cone types can be represented (according to a simple model of so-called opponent process theory presented in Boynton (1979)) by the following two functions.

(1) $C1\ ((b, g, r)) = r - g$
(2) $C2\ ((b, g, r)) = g + r - b$

It is assumed that the amount of greenness, yellowness, blueness and redness experienced by an observer in a concrete case can be predicted on the basis of the values of C1 and C2. In case $C1\ ((b, g, r)) = O$, the corresponding object will appear neither greenish nor reddish to the person. At the zero-point of the second function, there will be no blue or yellow component in the perceived colour. For positive values of C1, the person does not sense any greenness and the amount of redness increases with the distance from zero. With negative values of C1, the person does not sense any redness, and the amount of greenness increases with the distance from zero. Analogously yellow-sensations are correlated with values of C2 greater than zero and blueness-sensations with values of C2 smaller than zero.

According to the prevailing theory about red-green blind vision these people differ from normal people in the following respect: their G-cones and R-cones contain the same photopigment. Therefore the average stimulation of their R- and G-cones will be equal for any light stimulus. The value of C1 consequently will always be zero and it follows from the theory that nothing will appear reddish or greenish to the subject. One group of red-green-blind people (so-called protanopes) have the photopigment normally contained in the G-cones not only in their G-cones, but also in their R-cones. For the other group of red-green blind people the reverse is true: their G-cones and their R-cones both contain the photopigment normally contained in R-cones. According to a widely accepted model of the inheritance of colour vision defects, both genes, the one that causes production of the G-cone photopigment in R-cones and the one that causes production of the R-cone photopigment in G-cones, may be active simultaneously in one single individual. In these cases the photopigments of the two cone types at issue are simply exchanged. The result should be a person which does not have any obvious colour vision defect. These people are called pseudonormal since they appear to be normal but really are not. To any light stimulus their R-cones react like normally filled G-cones and their G-cones react like normally filled R-cones. The reversed filling of cones with photopigments only affects the causal interconnections between external stimuli and cone type activation. It does not, however, affect the causal interconnections between cone type activation and the states of the two chromatic channels. This second causal dependency is therefore assumed not to be altered in pseudonormal people. It follows that any light stimulus which causes the r-g-channel of a normal person to have the value y, will cause the r-g channel of a pseudonormal person to have the value $-y$. If y corresponds to a reddish component in the perceived colour, then $-y$ corresponds to a greenish component in the perceived colour (and vice versa). It therefore follows from received scientific theory about human colour vision, that pseudonormal people, if they exist, are red-green-inverted in the following sense: things that appear reddish to normal people to a certain degree, appear greenish to pseudonormal people to roughly the

same degree (and vice versa) while the perception of yellowish or bluish components remains unaffected.[4]

3. Philosophical Consequences

3.1. A Problem for Wittgensteinians

Let us call an N-case a case where a person P is red-green-inverted and yet there is no behavioural difference between P and normally sighted people detectable in normal life situations that would give any reason to suspect that P's colour perceptions differ from those of normal people. Some of those philosophers who are influenced by Wittgensteinian ideas think that the possibility of an N-case can be excluded without empirical research on the basis of philosophical considerations alone. They would subscribe to the following view: Ripe tomatoes look red to a given person if it is appropriate according to the rules of the relevant language game to assert that they look red to the person at issue. These rules do not require physiological examination of someone's visual system. Pseudonormal people are expected by colour vision science to behave roughly like normal people do in colour discrimination and colour judgement and therefore the conditions meant by Wittgensteinians for an appropriate ascription of normal colour perception are certainly fulfilled. So it seems that the Wittgensteinian must deny that pseudonormal people are red-green inverted and finds himself in conflict with what colour vision science asserts.

The Wittgensteinian however might defend his view claiming that the rules governing the use of colour appearance concepts in normal language are different from those governing scientific usage of these terms. He might then adopt one of the two following slightly different strategies: a) he might say that philosophy is concerned with everyday language and therefore need not care about how colour vision science describes the phenomenology of pseudonormal vision or b) he might admit that given the results of colour vision science sketched above we have reason to change the rules of the game and adopt the view that pseudonormal people are red-green inverted. In order to argue against both defence strategies it is necessary to show that colour vision science when using colour appearance terms does not introduce new concepts but rather uses these terms in their normal way. This indeed seems quite obvious, but to argue for this claim is a more complicated task which cannot be completed in the present paper.[5]

3.2. Pseudonormal Vision and Functionalism

It has been objected to functionalism that there could be what I will call an F-case. An F-case is a case where there is no relevant functional difference between a person P and normally sighted people although P is red-green-inverted.[6] Before we can begin discussing whether pseudonormals represent an F-case we need to distinguish different senses of 'functional difference' and thereby different versions of functionalism.

Conceptual functionalism claims that the meaning of mental terms may be analysed in functionalist terminology. According to conceptual functionalism to see something as red means to be in a state which plays a specific causal role. This causal role, according to conceptual functionalism, can be specified by reference to a) typical causes of the state and b) typical causal influence of the state at issue upon other mental states. The proponent of conceptual functionalism therefore must deny the possibility of what I will call an F1-case: An F1-case would be realized if for a red-green-inverted person P something like the following two conditions hold: a) P does not differ from normals with respect to colour naming and colour discrimination behaviour and b) if there is a specific difference in the role red- and green-sensations play in connection with emotions, other modes of perception, space perception and the like, then these roles in the case of person P are reversed too. For N cases it was required that the difference between P and normally sighted people could not be detected in normal life. It is required in addition for F1-cases that P will behave like a normal person even in sophisticated psychological and psychophysiological experiments. To the proponent of conceptual functionalism, we may ascribe the view that F1-cases are incoherent. To reject conceptual functionalism it is not necessary to show that pseudonormal people represent F1-cases. It suffices to argue that according to received colour vision science the question whether they do represent F1-cases or whether they do not need to be settled by empirical research. This is enough since no hypotheses seriously considered in scientific theory should be regarded incoherent by any philosophical proposal. It

has already been shown that pseudonormals, if they exist, are red-green inverted according to scientific theory. Whether they do represent an F1-case therefore only depends on the answer to the following question: Are there differences between red-sensations and green-sensations with respect to their causal influence upon other mental states which are innate and will not be overridden by learning processes? If the answer is 'yes,' then pseudonormal people cannot represent F1-cases and could be detected by sophisticated psychological experiments. If the answer is 'no,' then pseudonormal people could not be detected without direct investigation of their retina and they would represent F1-cases. Obviously the question needs to be settled by empirical research, and conceptual functionalism thus violates the above formulated prima facie constraint.[7]

Psychofunctionalism does not claim to give an analysis of the meaning of mental terms. Rather it proposes to accept the *empirical* hypothesis that mental terms will turn out to refer to functionally definable internal states. Colour vision science asserts that red sensations occur when the relevant r-g channel is in a specific type of state which is represented by positive values of C1. Let us call this type of state 'positive r-g channel state,' analogously I will talk of negative r-g channel states. The question of whether psychofunctionalism violates the above formulated prima facie constraint then depends on whether the difference between positive and negative r-g channel states is a functional difference in the sense of psychofunctionalism. At first sight it seems that it is not: positive and negative r-g channel states can be distinguished by reference to their causes. Positive states are caused by a predominance of R-cone activity, while negative states are caused by a predominance of G-cone activity. But this is a functionally describable difference only if the two types of cones can be functionally defined. The most obvious way to define receptor types in the present context is by reference to the way they react to light stimuli. This strategy however is not available here. R-cones can be filled with the photopigment normally contained in G-cones and thereby be caused to behave like G-cones. But, as the hypothesis of pseudonormal vision shows, colour vision science explicitly denies that a G-cone filled with the wrong pigment thereby ceases to be a G-cone. Of course there is a difference in causal role between G-cones and R-cones: They have different influences upon the channel states. But this is what we started with. So the difference between positive and negative r-g channel states cannot be functionally specified by reference to the way these states are caused. Still, the psychofunctionalist may hope that the two channel states will turn out to play different functional roles on higher levels of information processing.

I have characterized psychofunctionalism by the empirical hypothesis that terms for mental states will turn out to refer to functionally definable states. A somehow stronger claim is however in the spirit of psychofunctionalism. Those who subscribe to some kind of psychofunctionalism certainly would have expected any theory of colour information processing to be a functional theory from the outset. This would mean that the central notions of colour vision science at any of its historical stages should be explicable in functionalist terminology. This stronger claim, however, is quite obviously wrong.[8]

3.3. Fixing the Reference of Physiological Concepts

The *real* story about the development of colour vision science seems to be this: It is a central assumption of colour vision science which has been accepted from the very beginning of this empirical discipline and has turned out fruitful that for any of the four phenomenally basic hues there must be some specific physiological process responsible for the occurrence of that colour sensation. (The assumption is hold true for the whole range of sighted people independently of their specific kind of—normal or abnormal—colour vision). One first step in the development of modern colour vision science was to *postulate* the existence of one type of physiological process responsible for every basic hue sensation and to assume that any of these four processes allows for degrees which are correlated with the corresponding amount of f-ness (where f is a basic hue) in the phenomenally given colour. *Thus the reference of physiological concepts was fixed in colour vision science by definite descriptions formulated using phenomenal concepts* (e.g. "the process p such that the 'degree of p' is correlated with the amount of redness"). It was assumed as a working hypothesis that these phenomenal descriptions are successful in picking out specific physiological types.

If this description is correct, then phenomenal concepts used in their everyday meaning did play and still do play an essential role in the development of scientific terminology. If this is true, then the psychofunctionalist who wishes to uphold what I called his stronger claim, needs to show that our phenomenal concepts really are functional concepts. He thus has to support, in addition, some kind of conceptual functionalism. Conceptual functionalist, however, has already been shown to be inadequate.[9]

4. Two Objections

Here is a possible objection that needs to be discussed: Someone might propose to redefine R-cones, G-cones and B-cones in terms of their corresponding spectral sensitivity curves. This indeed would cause the argument to break down. We then could not say of pseudonormal people that their cones contain the 'wrong' pigment, since, by containing erythrolabe and thereby a specific spectral sensitivity curve, a receptor *by definition* becomes an R-cone. This definition of cone-types, combined with a definition of the relevant states of the r-g-channel according to its causal relations to the cone-types, leads to the conclusion that in normal subjects and in pseudonormal subjects the same external conditions cause the same r-g-states. It would follow that red things appear red to pseudonormal people just as they do to normal ones. The philosopher proposing this redefinition might make his view still more difficult to attack by adding: My position does not need the assumption that the proposed definition is more adequate than a morphological individuation of cone types. It probably is a matter of practical convenience which definition should be preferred. Since it depends on what definition we choose whether opponent process theory predicts normal vision or inverted vision for pseudonormal subjects, the question whether an object appears red or green to a pseudonormal person turns out to be decidable by convention. It is then not a factual question about what really is the case. This result—the opponent might go on—is almost as good or even better than genuine impossibility of qualia inversion.

This counterargument can be met in two ways: First, redefining receptor types in the way proposed would by definition exclude specific cases of *acquired* red-green-inversion which seems quite unacceptable.[10] Second, the proposal violates the widely accepted principle of supervenience for mental properties upon the relevant physiological properties: Let us for the moment accept that the relevant states of the r-g-channel can be functionally defined in the way proposed, and let us call states represented by positive C1-values, positive states of the r-g-channel and the same for negative values. Since the neural hardware is not affected by exchanging photopigments, we must assume that the physiological state produced by a specific pattern of stimulation of concrete photoreceptors in a given person is the same regardless of whether the photopigments are reversed. The proposal, therefore, entails that the same physiological state that realizes a positive r-g-state, given normal distribution of photopigments, realizes a negative r-g-state, given pseudonormal distribution of photopigments. So the proposed definition, combined with opponent process theory, entails the prediction that the *very same* physiological state will lead to a red-sensation in the one case and to a green-sensation in the other. Since the only difference between the two cases lies in the way the physiological state is *caused* (by different patterns of light stimuli) and since the brain does not have any access to this information, this would seem rather mysterious.

David Lewis defended functionalism against the so-called Inverted Qualia Argument claiming that "object o looks red to person x" is ambiguous, needs to be relativized to a population P and means something like the following: "object o produces in x a state, which in people of population P plays the role of red-perceptions" where 'the role of red-perceptions' is assumed to be explicable in functional terms.[11] On this account the assumption that green things look red to pseudonormal people (in the sense in which it is true) would *mean:* when looking at grass the brain of pseudonormal people is in a physiological state which occupies the role of seeing something red in normally sighted people.[12] Lewis' proposal, however, yields an inadequate interpretation of the following central assumption in colour vision science: there is a specific physiological state which is responsible for red sensations in general (whoever is in that state has an experience of red and vice versa). This assumption is supposed to be true for all human beings in a non-trivial sense. This basic assumption, therefore, should not follow from the following 'weaker' claim: There is a

physiological state (or process) which occupies a specific functional role F in normal subjects. On Lewis' account, however, it does.[13] This argument does not in itself show that the proposal violates the above formulated constraint for philosophical theories, but it does prove the violation of another plausible necessary condition for an adequate philosophical theory: *If a hypothesis H which is accepted or seriously considered in some well-established scientific theory contains a concept C and if the philosophical theory proposes a definition of C, then replacing C by the proposed definiens should not change the empirical content of H.*

5. Final Remark

The two constraints used in this paper only provide prima facie reasons for rejecting a given philosophical proposal. They may be overridden by philosophical considerations in some cases even if the scientific theory is empirically well-established. However, in such a case, the philosopher who wishes to reject scientific terminology, should be able to argue convincingly that the theory can be replaced by an alternative one, which does conform to the philosopher's intuitions and is yet in some relevant sense empirically equivalent to the original one.[14]

NOTES

1. Sydney Shoemaker, "The Inverted Spectrum," *The Journal of Philosophy* 79 (1982): 357–82, p. 364.
2. See T. P. Piantanida, "A replacement model of X-linked recessive colour vision defects," *Annals of Human Genetics* 37 (1974): 393–404; and Robert M. Boynton, *Human Color Vision*, (New York Holt, Rinehart & Winston, 1979), p. 351–58.
3. Boynton in *Human Color Vision*, op. cit., p. 356.
4. In an earlier paper and in my dissertation I discussed the case of photopigment exchange between R- and G-cones as an empirically possible but only imaginary case for which colour vision science would have to predict red-green inverted vision (see my "Irreduzibel mentale Prädikate in physiologischen Theorien der Farbwahrnehmung," *Berichte des Internationalen Wittgensteinsymposiums* 1988, Wien 1989 S. 59–62 and my "Farben und phänomenales Wissen," Conceptus Studienband 9, Wien 1993: Academia Verlag (St. Augustin)). Three years later I discovered that the imaginary case had actually appeared as a serious hypothesis in scientific literature. As far as I know pseudonormal vision has not yet been discussed in philosophical literature which might be due to the fact that the hypothesis can only be found in chapters or articles about the inheritance of colour vision deficiencies which philosophers might tend to skip.
5. If pseudonormal people exist, then normal subjects are systematically wrong about the colour experiences of these people as long as they believe them to be normal. Both strategies discussed above would commit the Wittgensteinian to the view that prior to the development of modern colour vision science there was no such error.
6. See, e.g., Ned Block and Jerry Fodor, "What Mental States Are Not," *Philosophical Review* 81 (1972): pp. 158–82, pp. 172–74.
7. It violates the constraint in the following way: a hypothesis which according to colour vision science needs to be settled by empirical research (the hypothesis that there are F1-cases) is incoherent according to conceptual functionalism. The case shows that conceptual functionalism violates a further plausible prima facie constraint: No claim should be conceptually true according to a philosophical theory if it has to be settled by empirical research according to colour vision science. (Conceptual functionalism violates this further constraint with respect to the hypothesis that there are no F1-cases.)

   The argument against conceptual functionalism may also be put this way: According to received scientific theory (according to central claims of colour vision science plus the hypothesis of pseudonormal vision) there are F1-cases if there are no innate differences in our reactions to red and green. Our colour concepts do not suffice to tell us that there are such innate differences. Therefore, contrary to conceptual functionalism, the existence of F1-cases cannot be excluded by conceptual considerations either. (I am grateful for a comment by Ned Block, which brought me to see this alternative way of making this point.)
8. It has been pointed out to me independently by Janet Levin and by Ned Block that the functionalist might reply claiming that normals and pseudonormals looking at a red thing simply have different *physiological realizations* of the same sensory quality. Of course, this is what some functionalists would like to say about the case. My point is that this description of pseudonormal vision (we would have to say, e.g., that red things look red to them) is in conflict with the way the case is described in color vision science. This reply, therefore, does not meet the above formulated prima facie constraint for philosophical theories: According to this account certain hypotheses accepted in colour vision science turn out false.
9. It is just a historical observation which is not in need of any philosophical argument that phenomenal concepts ('sensations of blue,' 'sensations of yellow' etc.) were used to pick out physiological types in the way roughly described in the text. For my argument I do not need the stronger claim that these concepts as used in these contexts cannot be interpreted in a behaviourist or functionalist manner (although I am certain they cannot). My point is, rather, that given the above historical observation, the stronger claim of the psychofunctionalist fails unless it is combined with some kind of conceptual functionalism.
10. Assume that someone's R-cones and G-cones start to produce the wrong photopigments at some point in his adult life. Colour vision science predicts

such a person will experience and report a radical change in his colour perception. Who accepts the proposed redefinition of cone types and subscribes to opponent process theory, however, would have to insist that no such change has taken place: Those individual receptors that were R-cones before the inversion of photopigment distribution in the retina of the person at issue, *turned into G-cones* according to the proposed redefinition. Thus, green objects cause a predominance of G-cone-activity before the inversion *and* after the inversion. Therefore, the channel state produced by green things is a negative r-g-channel state before *and* after the change. So, according to the proposed redefinition, acquired photopigment inversion could not result in any change in the colour perceived by the subject.

11. See David Lewis, "Mad Pain and Martian Pain," in *Philosophy of Psychology*, vol. I, Ned Block, ed., (Cambridge: Harvard University Press, 1980), pp. 216–22, p. 200.
12. Assuming that pseudonormal and normal people are functionally equivalent in the relevant sense, on Lewis' account the following further assumptions hold: green things look green to pseudonormal people relative to the group of pseudonormal people, green things look red to normal people relative to the group of pseudonormal people, green things look green to normal people relative to their own group. Lewis's proposal, of course, should not be confused with the view that pseudonormal people and normal people simply refer to different subjective qualities when they use colour appearance concepts.
13. The argument can be formulated more precisely:

    (A) $\exists s \forall x (\langle s,x\rangle \in \alpha \leftrightarrow R(x))$
    (B) $\exists s \forall x (\langle s,x\rangle \in \alpha \leftrightarrow \langle \text{the } s' FR(s',P^*),x\rangle \in \alpha)$
    (C) there is exactly one s such that FR(s,P*)

$\langle s,x\rangle \in \alpha$	: the brain of the person x is in the physiological state s
R(x)	: x has a sensation of red
FR(s,P)	: the state s occupies the functional role of seeing something as red in population P
P*	: population of normally sighted people
The s ϕ[s]	: the state s which satisfies ϕ. Quantifiers followed by s or s′ quantify over physiological states, quantifiers followed by x quantify over people.

    (A) is the basic assumption at issue. (B) is the account of Lewis for this assumption (R(x) is replaced by the proposed definiens). (B), however, logically follows from (C) and therefore cannot be equivalent to (A) as meant in colour vision science.
14. I have benefitted from discussions on this topic with Max Drömmer, Andreas Kemmerling, Martin Rechenauer, and Wolfgang Spohn. I am very grateful to Ned Block for detailed criticisms of an earlier version of the paper. The work was supported by the grant Nr. Sp. 279/4-1 from the *Deutsche Forschungsgemeinschaft*. Special thanks is due to Edith Vanghelof who helped with linguistic corrections.

Mental Events

Donald Davidson

Mental events such as perceivings, rememberings, decisions, and actions resist capture in the nomological net of physical theory.[1] How can this fact be reconciled with the causal role of mental events in the physical world? Reconciling freedom with causal determinism is a special case of the problem if we suppose that causal determinism entails capture in, and freedom requires escape from, the nomological net. But the broader issue can remain alive even for someone who believes a correct analysis of free action reveals no conflict with determinism. *Autonomy* (freedom, self-rule) may or may not clash with determinism; *anomaly* (failure to fall under a law) is, it would seem, another matter.

I start from the assumption that both the causal dependence, and the anomalousness, of mental events are undeniable facts. My aim is therefore to explain, in the face of apparent difficulties, how this can be. I am in sympathy with Kant when he says,

> it is as impossible for the subtlest philosophy as for the commonest reasoning to argue freedom away. Philosophy must therefore assume that no true contradiction will be found between freedom and natural necessity in the same human actions, for it cannot give up the idea of nature any more than that of freedom. Hence even if we should never be able to conceive how freedom is possible, at least this apparent contradiction must be convincingly eradicated. For if the thought of freedom contradicts itself or nature . . . it would have to be surrendered in competition with natural necessity.[2]

Generalize human actions to mental events, substitute anomaly for freedom, and this is a description of my problem. And of course the connection is closer, since Kant believed freedom entails anomaly.

Now let me try to formulate a little more carefully the 'apparent contradiction' about mental events that I want to discuss and finally dissipate. It may be seen as stemming from three principles.

The first principle asserts that at least some mental events interact causally with physical events. (We could call this the Principle of Causal Interaction.) Thus for example if someone sank the *Bismarck*, then various mental events such as perceivings, notings, calculations, judgments, decisions, intentional actions, and changes of belief played a causal role in the sinking of the *Bismarck.* In particular, I would urge that the fact that someone sank the *Bismarck* entails that he moved his body in a way that was caused by mental events of certain sorts, and that this bodily movement in turn caused the *Bismarck* to sink.[3] Perception illustrates how causality may run from the physical to the mental: if a man perceives that a ship is approaching, then a ship approaching must have caused him to come to believe that a ship is approaching. (Nothing depends on accepting these as examples of causal interaction.)

Though perception and action provide the most obvious cases where mental and physical events interact causally, I think reasons could be given for the view that all mental events ultimately, perhaps through causal relations with other mental events, have causal intercourse with physical events. But if there are mental events that have no physical events as causes or effects, the argument will not touch them.

The second principle is that where there is causality, there must be a law: events related as cause and effect fall under strict deterministic laws. (We may term this the Principle of the Nomological Character of Causality.) This principle, like the first, will be treated here as

From L. Foster and J. Swanson, eds., *Experience and Theory* (Humanities Press, 1970), pp. 79–101. Reprinted with permission of the author.

an assumption, though I shall say something by way of interpretation.[4]

The third principle is that there are no strict deterministic laws on the basis of which mental events can be predicted and explained (the Anomalism of the Mental).

The paradox I wish to discuss arises for someone who is inclined to accept these three assumptions or principles, and who thinks they are inconsistent with one another. The inconsistency is not, of course, formal unless more premises are added. Nevertheless it is natural to reason that the first two principles, that of causal interaction, and that of the nomological character of causality, together imply that at least some mental events can be predicted and explained on the basis of laws, while the principle of the anomalism of the mental denies this. Many philosophers have accepted, with or without argument, the view that the three principles do lead to a contradiction. It seems to me, however, that all three principles are true, so that what must be done is to explain away the appearance of contradiction; essentially the Kantian line.

The rest of this paper falls into three parts. The first part describes a version of the identity theory of the mental and the physical that shows how the three principles may be reconciled. The second part argues that there cannot be strict psychophysical laws; this is not quite the principle of the anomalism of the mental, but on reasonable assumptions entails it. The last part tries to show that from the fact that there can be no strict psychophysical laws, and our other two principles, we can infer the truth of a version of the identity theory, that is, a theory that identifies at least some mental events with physical events. It is clear that this 'proof' of the identity theory will be at best conditional, since two of its premises are unsupported, and the argument for the third may be found less than conclusive. But even someone unpersuaded of the truth of the premises may be interested to learn how they may be reconciled and that they serve to establish a version of the identity theory of the mental. Finally, if the argument is a good one, it should lay to rest the view, common to many friends and some foes of identity theories, that support for such theories can come only from the discovery of psychophysical laws.

I

The three principles will be shown consistent with one another by describing a view of the mental and the physical that contains no inner contradiction and that entails the three principles. According to this view, mental events are identical with physical events. Events are taken to be unrepeatable, dated individuals such as the particular eruption of a volcano, the (first) birth or death of a person, the playing of the 1968 World Series, or the historic utterance of the words, 'You may fire when ready, Gridley.' We can easily frame identity statements about individual events; examples (true or false) might be:

The death of Scott = the death of the author of *Waverley;*

The assassination of the Archduke Ferdinand = the event that started the First World War;

The eruption of Vesuvius in A.D. 79 = the cause of the destruction of Pompeii.

The theory under discussion is silent about processes, states, and attributes if these differ from individual events.

What does it mean to say that an event is mental or physical? One natural answer is that an event is physical if it is describable in a purely physical vocabulary, mental if describable in mental terms. But if this is taken to suggest that an event is physical, say, if some physical predicate is true of it, then there is the following difficulty. Assume that the predicate '*x* took place at Noosa Heads' belongs to the physical vocabulary; then so also must the predicate '*x* did not take place at Noosa Heads' belong to the physical vocabulary. But the predicate '*x* did or did not take place at Noosa Heads' is true of every event, whether mental or physical.[5] We might rule out predicates that are tautologically true of every event, but this will not help since every event is truly describable either by '*x* took place at Noosa Heads' or by '*x* did not take place at Noosa Heads.' A different approach is needed.[6]

We may call those verbs mental that express propositional attitudes like believing, intending, desiring, hoping, knowing, perceiving, noticing, remembering, and so on. Such verbs are characterized by the fact that they sometimes feature in sentences with subjects that refer to persons, and are completed by embedded sentences in which the usual rules of substitution appear to break down. This criterion is not precise, since I do not want to include these verbs when they occur in contexts that are fully extensional ('He knows Paris,' 'He perceives the moon' may be cases), nor exclude them

whenever they are not followed by embedded sentences. An alternative characterization of the desired class of mental verbs might be that they are psychological verbs as used when they create apparently nonextensional contexts.

Let us call a description of the form 'the event that is M' or an open sentence of the form 'event x is M' a *mental description* or a *mental open sentence* if and only if the expression that replaces 'M' contains at least one mental verb essentially. (Essentially, so as to rule out cases where the description or open sentence is logically equivalent to one not containing mental vocabulary.) Now we may say that an event is mental if and only if it has a mental description, or (the description operator not being primitive) if there is a mental open sentence true of that event alone. Physical events are those picked out by descriptions or open sentences that contain only the physical vocabulary essentially. It is less important to characterize a physical vocabulary because relative to the mental it is, so to speak, recessive in determining whether a description is mental or physical. (There will be some comments presently on the nature of a physical vocabulary, but these comments will fall far short of providing a criterion.)

On the proposed test of the mental, the distinguishing feature of the mental is not that it is private, subjective, or immaterial, but that it exhibits what Brentano called intentionality. Thus intentional actions are clearly included in the realm of the mental along with thoughts, hopes, and regrets (or the events tied to these). What may seem doubtful is whether the criterion will include events that have often been considered paradigmatic of the mental. Is it obvious, for example, that feeling a pain or seeing an afterimage will count as mental? Sentences that report such events seem free from taint of nonextensionality, and the same should be true of reports of raw feels, sense data, and other uninterpreted sensations, if there are any.

However, the criterion actually covers not only the havings of pains and afterimages, but much more besides. Take some event one would intuitively accept as physical, let's say the collision of two stars in distant space. There must be a purely physical predicate 'Px' true of this collision, and of others, but true of only this one at the time it occurred. This particular time, though, may be pinpointed as the same time that Jones notices that a pencil starts to roll across his desk. The distant stellar collision is thus *the* event x such that Px and x is simultaneous with Jones' noticing that a pencil starts to roll across his desk. The collision has now been picked out by a mental description and must be counted as a mental event.

This strategy will probably work to show every event to be mental; we have obviously failed to capture the intuitive concept of the mental. It would be instructive to try to mend this trouble, but it is not necessary for present purposes. We can afford Spinozistic extravagance with the mental since accidental inclusions can only strengthen the hypothesis that all mental events are identical with physical events. What would matter would be failure to include bona fide mental events, but of this there seems to be no danger.

I want to describe, and presently to argue for, a version of the identity theory that denies that there can be strict laws connecting the mental and the physical. The very possibility of such a theory is easily obscured by the way in which identity theories are commonly defended and attacked. Charles Taylor, for example, agrees with protagonists of identity theories that the sole 'ground' for accepting such theories is the supposition that correlations or laws can be established linking events described as mental with events described as physical. He says, 'It is easy to see why this is so: unless a given mental event is invariably accompanied by a given, say, brain process, there is no ground for even mooting a general identity between the two.'[7] Taylor goes on (correctly, I think) to allow that there may be identity without correlating laws, but my present interest is in noticing the invitation to confusion in the statement just quoted. What can 'a given mental event' mean here? Not a particular, dated, event, for it would not make sense to speak of an individual event being 'invariably accompanied' by another. Taylor is evidently thinking of events of a given *kind*. But if the only identities are of kinds of events, the identity theory presupposes correlating laws.

One finds the same tendency to build laws into the statement of the identity theory in these typical remarks:

> When I say that a sensation is a brain process or that lightning is an electrical discharge, I am using 'is' in the sense of strict identity . . . there are not two things: a flash of lightning and an electrical discharge. There is one thing, a flash of lightning, which is described scientifically as an electrical discharge to the earth from a cloud of ionized water molecules.[8]

The last sentence of this quotation is perhaps to be understood as saying that for every lightning

flash there exists an electrical discharge to the earth from a cloud of ionized water molecules with which it is identical. Here we have an honest ontology of individual events and can make literal sense of identity. We can also see how there could be identities without correlating laws. It is possible, however, to have an ontology of events with the conditions of individuation specified in such a way that any identity implies a correlating law. Kim, for example, suggests that F*a* and G*b* 'describe or refer to the same event' if and only if $a = b$ and the property of being F = the property of being G. The identity of the properties in turn entails that (x) $(Fx \leftrightarrow Gx)$.[9] No wonder Kim says:

> If pain is identical with brain state *B*, there must be a concomitance between occurrences of pain and occurrences of brain state *B*. . . . Thus, a necessary condition of the pain–brain state *B* identity is that the two expressions 'being in pain' and 'being in brain state *B*' have the same extension. . . . There is no conceivable observation that would confirm or refute the identity but not the associated correlation.[10]

It may make the situation clearer to give a fourfold classification of theories of the relation between mental and physical events that emphasizes the independence of claims about laws and claims of identity. On the one hand there are those who assert, and those who deny, the existence of psychophysical laws; on the other hand there are those who say mental events are identical with physical and those who deny this. Theories are thus divided into four sorts: *Nomological monism*, which affirms that there are correlating laws and that the events correlated are one (materialists belong in this category); *nomological dualism*, which comprises various forms of parallelism, interactionism, and epiphenomenalism; *anomalous dualism*, which combines onto-logical dualism with the general failure of laws correlating the mental and the physical (Cartesianism). And finally there is *anomalous monism*, which classifies the position I wish to occupy.[11]

Anomalous monism resembles materialism in its claim that all events are physical, but rejects the thesis, usually considered essential to materialism, that mental phenomena can be given purely physical explanations. Anomalous monism shows an ontological bias only in that it allows the possibility that not all events are mental, while insisting that all events are physical. Such a bland monism, unbuttressed by correlating laws or conceptual economies, does not seem to merit the term 'reductionism'; in any case it is not apt to inspire the nothing-but reflex ('Conceiving the *Art of the Fugue* was nothing but a complex neural event,' and so forth.)

Although the position I describe denies there are psychophysical laws, it is consistent with the view that mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect. Dependence or supervenience of this kind does not entail reducibility through law or definition: if it did, we could reduce moral properties to descriptive, and this there is good reason to *believe* cannot be done; and we might be able to reduce truth in a formal system to syntactical properties, and this we *know* cannot in general be done.

This last example is in useful analogy with the sort of lawless monism under consideration. Think of the physical vocabulary as the entire vocabulary of some language L with resources adequate to express a certain amount of mathematics, and its own syntax. L′ is L augmented with the truth predicate 'true-in-L,' which is 'mental.' In L (and hence L′) it is possible to pick out, with a definite description or open sentence, each sentence in the extension of the truth predicate, but if L is consistent there exists no predicate of syntax (of the 'physical' vocabulary), no matter how complex, that applies to all and only the true sentences of L. There can be no 'psychophysical law' in the form of a biconditional, '(x) (x is true-in-L if and only if x is ϕ)' where 'ϕ' is replaced by a 'physical' predicate (a predicate of L). Similarly, we can pick out each mental event using the physical vocabulary alone, but no purely physical predicate, no matter how complex, has, as a matter of law, the same extension as a mental predicate.

It should now be evident how anomalous monism reconciles the three original principles. Causality and identity are relations between individual events no matter how described. But laws are linguistic; and so events can instantiate laws, and hence be explained or predicted in the light of laws, only as those events are described in one or another way. The principle of causal interaction deals with events in extension and is therefore blind to the mental–physical dichotomy. The principle of the anomalism of the mental concerns events described as mental, for events are mental only as described. The principle of the nomological character of causality

must be read carefully: it says that when events are related as cause and effect, they have descriptions that instantiate a law. It does not say that every true singular statement of causality instantiates a law.[12]

II

The analogy just bruited, between the place of the mental amid the physical, and the place of the semantical in a world of syntax, should not be strained. Tarski proved that a consistent language cannot (under some natural assumptions) contain an open sentence 'Fx' true of all and only the true sentences of that language. If our analogy were pressed, then we would expect a proof that there can be no physical open sentence 'Px' true of all and only the events having some mental property. In fact, however, nothing I can say about the irreducibility of the mental deserves to be called a proof; and the kind of irreducibility is different. For if anomalous monism is correct, not only can every mental event be uniquely singled out using only physical concepts, but since the number of events that falls under each mental predicate may, for all we know, be finite, there may well exist a physical open sentence coextensive with each mental predicate, though to construct it might involve the tedium of a lengthy and uninstructive alternation. Indeed, even if finitude is not assumed, there seems no compelling reason to deny that there could be coextensive predicates, one mental and one physical.

The thesis is rather that the mental is nomologically irreducible: there may be *true* general statements relating the mental and the physical, statements that have the logical form of a law; but they are not *lawlike* (in a strong sense to be described). If by absurdly remote chance we were to stumble on a nonstochastic true psychophysical generalization, we would have no reason to believe it more than roughly true.

Do we, by declaring that there are no (strict) psychophysical laws, poach on the empirical preserves of science—a form of hubris against which philosophers are often warned? Of course, to judge a statement lawlike or illegal is not to decide its truth outright; relative to the acceptance of a general statement on the basis of instances, ruling it lawlike must be a priori. But such relative apriorism does not in itself justify philosophy, for in general the grounds for deciding to trust a statement on the basis of its instances will in turn be governed by theoretical and empirical concerns not to be distinguished from those of science. If the case of supposed laws linking the mental and the physical is different, it can only be because to allow the possibility of such laws would amount to changing the subject. By changing the subject I mean here: deciding not to accept the criterion of the mental in terms of the vocabulary of the propositional attitudes. This short answer cannot prevent further ramifications of the problem, however, for there is no clear line between changing the subject and changing what one says on an old subject, which is to admit, in the present context at least, that there is no clear line between philosophy and science. Where there are no fixed boundaries only the timid never risk trespass.

It will sharpen our appreciation of the anomological character of mental—physical generalizations to consider a related matter, the failure of definitional behaviorism. Why are we willing (as I assume we are) to abandon the attempt to give explicit definitions of mental concepts in terms of behavioral ones? Not, surely, just because all actual tries are conspicuously inadequate. Rather it is because we are persuaded, as we are in the case of so many other forms of definitional reductionism (naturalism in ethics, instrumentalism and operationalism in the sciences, the causal theory of meaning, phenomenalism, and so on—the catalogue of philosophy's defeats), that there is system in the failures. Suppose we try to say, not using any mental concepts, what it is for a man to believe there is life on Mars. One line we could take is this: when a certain sound is produced in the man's presence ('Is there life on Mars?') he produces another ('Yes'). But of course this shows he believes there is life on Mars only if he understands English, his production of the sound was intentional, and was a response to the sounds as meaning something in English; and so on. For each discovered deficiency, we add a new proviso. Yet no matter how we patch and fit the nonmental conditions, we always find the need for an additional condition (provided he *notices, understands*, etc.) that is mental in character.[13]

A striking feature of attempts at definitional reduction is how little seems to hinge on the question of synonymy between definiens and definiendum. Of course, by imagining counterexamples we do discredit claims of synonymy. But the pattern of failure prompts a stronger conclusion: if we were to find an open sentence couched in behavioral terms and exactly

coextensive with some mental predicate, nothing could reasonably persuade us that we had found it. We know too much about thought and behavior to trust exact and universal statements linking them. Beliefs and desires issue in behavior only as modified and mediated by further beliefs and desires, attitudes and attendings, without limit. Clearly this holism of the mental realm is a clue both to the autonomy and to the anomalous character of the mental.

These remarks apropos definitional behaviorism provide at best hints of why we should not expect nomological connections between the mental and the physical. The central case invites further consideration.

Lawlike statements are general statements that support counterfactual and subjunctive claims, and are supported by their instances. There is (in my view) no nonquestion-begging criterion of the lawlike, which is not to say there are no reasons in particular cases for a judgment. Lawlikeness is a matter of degree, which is not to deny that there may be cases beyond debate. And within limits set by the conditions of communication, there is room for much variation between individuals in the pattern of statements to which various degrees of nomologicality are assigned. In all these respects, nomologicality is much like analyticity, as one might expect since both are linked to meaning.

'All emeralds are green' is lawlike in that its instances confirm it, but 'all emeralds are grue' is not, for 'grue' means 'observed before time t and green, otherwise blue,' and if our observations were all made before t and uniformly revealed green emeralds, this would not be a reason to expect other emeralds to be blue. Nelson Goodman has suggested that this shows that some predicates, 'grue' for example, are unsuited to laws (and thus a criterion of suitable predicates could lead to a criterion of the lawlike). But it seems to me the anomalous character of 'All emeralds are grue' shows only that the predicates 'is an emerald' and 'is grue' are not suited to one another: grueness is not an inductive property of emeralds. Grueness *is* however an inductive property of entities of other sorts, for instance of emerires. (Something is an emerire if it is examined before t and is an emerald, and otherwise is a sapphire.) Not only is 'All emerires are grue' entailed by the conjunction of the lawlike statements 'All emeralds are green' and 'All sapphires are blue,' but there is no reason, as far as I can see, to reject the deliverance of intuition, that it is itself lawlike.[14] Nomological statements bring together predicates that we know a priori are made for each other—know, that is, independently of knowing whether the evidence supports a connection between them. 'Blue,' 'red,' and 'green' are made for emeralds, sapphires, and roses; 'grue,' 'bleen,' and 'gred' are made for sapphalds, emerires, and emeroses.

The direction in which the discussion seems headed is this: mental and physical predicates are not made for one another. In point of lawlikeness, psychophysical statements are more like 'All emeralds are grue' than like 'All emeralds are green.'

Before this claim is plausible, it must be seriously modified. The fact that emeralds examined before t are grue not only is no reason to believe all emeralds are grue; it is not even a reason (if we know the time) to believe *any* unobserved emeralds are grue. But if an event of a certain mental sort has usually been accompanied by an event of a certain physical sort, this often is a good reason to expect other cases to follow suit roughly in proportion. The generalizations that embody such practical wisdom are assumed to be only roughly true, or they are explicitly stated in probabilistic terms, or they are insulated from counterexample by generous escape clauses. Their importance lies mainly in the support they lend singular causal claims and related explanations of particular events. The support derives from the fact that such a generalization, however crude and vague, may provide good reason to believe that underlying the particular case there is a regularity that could be formulated sharply and without caveat.

In our daily traffic with events and actions that must be foreseen or understood, we perforce make use of the sketchy summary generalization, for we do not know a more accurate law, or if we do, we lack a description of the particular events in which we are interested that would show the relevance of the law. But there is an important distinction to be made within the category of the rude rule of thumb. On the one hand, there are generalizations whose positive instances give us reason to believe the generalization itself could be improved by adding further provisos and conditions stated in the same general vocabulary as the original generalization. Such a generalization points to the form and vocabulary of the finished law: we may say that it is a *homonomic* generalization. On the other hand there are generalizations which when instantiated may give us reason to believe there is a precise law at work, but one that can be stated only by shifting to a different

vocabulary. We may call such generalizations *heteronomic*.

I suppose most of our practical lore (and science) is heteronomic. This is because a law can hope to be precise, explicit, and as exceptionless as possible only if it draws its concepts from a comprehensive closed theory. This ideal theory may or may not be deterministic, but it is if any true theory is. Within the physical sciences we do find homonomic generalizations, generalizations such that if the evidence supports them, we then have reason to believe they may be sharpened indefinitely by drawing upon further physical concepts: there is a theoretical asymptote of perfect coherence with all the evidence, perfect predictability (under the terms of the system), total explanation (again under the terms of the system). Or perhaps the ultimate theory is probabilistic, and the asymptote is less than perfection; but in that case there will be no better to be had.

Confidence that a statement is homonomic, correctible within its own conceptual domain, demands that it draw its concepts from a theory with strong constitutive elements. Here is the simplest possible illustration; if the lesson carries, it will be obvious that the simplification could be mended.

The measurement of length, weight, temperature, or time depends (among many other things, of course) on the existence in each case of a two-place relation that is transitive and asymmetric: warmer than, later than, heavier than, and so forth. Let us take the relation *longer than* as our example. The law or postulate of transitivity is this:

(L) $L(x,y)$ and $L(y,z) \rightarrow L(x,z)$

Unless this law (or some sophisticated variant) holds, we cannot easily make sense of the concept of length. There will be no way of assigning numbers to register even so much as ranking in length, let alone the more powerful demands of measurement on a ratio scale. And this remark goes not only for any three items directly involved in an intransitivity: it is easy to show (given a few more assumptions essential to measurement of length) that there is no consistent assignment of a ranking to any item unless (L) holds in full generality.

Clearly (L) alone cannot exhaust the import of 'longer than'—otherwise it would not differ from 'warmer than' or 'later than.' We must suppose there is some empirical content, however difficult to formulate in the available vocabulary, that distinguishes 'longer than' from the other two-place transitive predicates of measurement and on the basis of which we may assert that one thing is longer than another. Imagine this empirical content to be partly given by the predicate '$O(x, y)$.' So we have this 'meaning postulate':

(M) $O(x,y) \rightarrow L(x,y)$

that partly interprets (L). But now (L) and (M) together yield an empirical theory of great strength, for together they entail that there do not exist three objects *a*, *b*, and *c* such that $O(a, b)$, $O(b, c)$, and $O(c, a)$. Yet what is to prevent this happening if '$O(x, y)$' is a predicate we can ever, with confidence, apply? Suppose we *think* we observe an intransitive triad; what do we say? We could count (L) false, but then we would have no application for the concept of length. We could say (M) gives a wrong test for length; but then it is unclear what we thought was the *content* of the idea of one thing being longer than another. Or we could say that the objects under observation are not, as the theory requires, *rigid* objects. It is a mistake to think we are forced to accept some one of these answers. Concepts such as that of length are sustained in equilibrium by a number of conceptual pressures, and theories of fundamental measurement are distorted if we force the decision, among such principles as (L) and (M): analytic or synthetic. It is better to say the whole set of axioms, laws, or postulates for the measurement of length is partly constitutive of the idea of a system of macroscopic, rigid, physical objects. I suggest that the existence of lawlike statements in physical science depends upon the existence of constitutive (or synthetic a priori) laws like those of the measurement of length within the same conceptual domain.

Just as we cannot intelligibly assign a length to any object unless a comprehensive theory holds of objects of that sort, we cannot intelligibly attribute any propositional attitude to an agent except within the framework of a viable theory of his beliefs, desires, intentions, and decisions.

There is no assigning beliefs to a person one by one on the basis of his verbal behavior, his choices, or other local signs no matter how plain and evident, for we make sense of particular beliefs only as they cohere with other beliefs, with preferences, with intentions, hopes, fears, expectations, and the rest. It is not merely, as with the measurement of length, that each case tests a theory and depends upon it, but that the content of a propositional attitude derives from its place in the pattern.

Crediting people with a large degree of consistency cannot be counted mere charity: it is unavoidable if we are to be in a position to accuse them meaningfully of error and some degree of irrationality. Global confusion, like universal mistake, is unthinkable, not because imagination boggles, but because too much confusion leaves nothing to be confused about and massive error erodes the background of true belief against which alone failure can be construed. To appreciate the limits to the kind and amount of blunder and bad thinking we can intelligibly pin on others is to see once more the inseparability of the question what concepts a person commands and the question what he does with those concepts in the way of belief, desire, and intention. To the extent that we fail to discover a coherent and plausible pattern in the attitudes and actions of others we simply forego the chance of treating them as persons.

The problem is not bypassed but given center stage by appeal to explicit speech behavior. For we could not begin to decode a man's sayings if we could not make out his attitudes towards his sentences, such as holding, wishing, or wanting them to be true. Beginning from these attitudes, we must work out a theory of what he means, thus simultaneously giving content to his attitudes and to his words. In our need to make him make sense, we will try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own lights, it goes without saying). Life being what it is, there will be no simple theory that fully meets these demands. Many theories will effect a more or less acceptable compromise, and between these theories there may be no objective grounds for choice.

The heteronomic character of general statements linking the mental and the physical traces back to this central role of translation in the description of all propositional attitudes, and to the indeterminacy of translation.[15] There are no strict psychophysical laws because of the disparate commitments of the mental and physical schemes. It is a feature of physical reality that physical change can be explained by laws that connect it with other changes and conditions physically described. It is a feature of the mental that the attribution of mental phenomena must be responsible to the background of reasons, beliefs, and intentions of the individual. There cannot be tight connections between the realms if each is to retain allegiance to its proper source of evidence. The nomological irreducibility of the mental does not derive merely from the seamless nature of the world of thought, preference and intention, for such interdependence is common to physical theory, and is compatible with there being a single right way of interpreting a man's attitudes without relativization to a scheme of translation. Nor is the irreducibility due simply to the possibility of many equally eligible schemes, for this is compatible with an arbitrary choice of one scheme relative to which assignments of mental traits are made. The point is rather that when we use the concepts of belief, desire and the rest, we must stand prepared, as the evidence accumulates, to adjust our theory in the light of considerations of overall cogency: the constitutive ideal of rationality partly controls each phase in the evolution of what must be an evolving theory. An arbitrary choice of translation scheme would preclude such opportunistic tempering of theory; put differently, a right arbitrary choice of a translation manual would be of a manual acceptable in the light of all possible evidence, and this is a choice we cannot make. We must conclude, I think, that nomological slack between the mental and the physical is essential as long as we conceive of man as a rational animal.

III

The gist of the foregoing discussion, as well as its conclusion, will be familiar. That there is a categorial difference between the mental and the physical is a commonplace. It may seem odd that I say nothing of the supposed privacy of the mental, or the special authority an agent has with respect to his own propositional attitudes, but this appearance of novelty would fade if we were to investigate in more detail the grounds for accepting a scheme of translation. The step from the categorial difference between the mental and the physical to the impossibility of strict laws relating them is less common, but certainly not new. If there is a surprise, then, it will be to find the lawlessness of the mental serving to help establish the identity of the mental with that paradigm of the lawlike, the physical.

The reasoning is this. We are assuming, under the Principle of the Causal Dependence of the Mental, that some mental events at least are causes or effects of physical events; the argument applies only to these. A second Principle (of the Nomological Character of Causality) says that each true singular causal statement

is backed by a strict law connecting events of kinds to which the events mentioned as cause and effect belong. Where there are rough, but homonomic, laws, there are laws drawing on concepts from the same conceptual domain and upon which there is no improving in point of precision and comprehensiveness. We urged in the last section that such laws occur in the physical sciences. Physical theory promises to provide a comprehensive closed system guaranteed to yield a standardized, unique description of every physical event couched in a vocabulary amenable to law.

It is not plausible that mental concepts alone can provide such a framework, simply because the mental does not, by our first principle, constitute a closed system. Too much happens to affect the mental that is not itself a systematic part of the mental. But if we combine this observation with the conclusion that no psychophysical statement is, or can be built into, a strict law, we have the Principle of the Anomalism of the Mental: there are no strict laws at all on the basis of which we can predict and explain mental phenomena.

The demonstration of identity follows easily. Suppose m, a mental event, caused p, a physical event; then under some description m and p instantiate a strict law. This law can only be physical, according to the previous paragraph. But if m falls under a physical law, it has a physical description; which is to say it is a physical event. An analogous argument works when a physical event causes a mental event. So every mental event that is causally related to a physical event is a physical event. In order to establish anomalous monism in full generality it would be sufficient to show that every mental event is cause or effect of some physical event; I shall not attempt this.

If one event causes another, there is a strict law which those events instantiate when properly described. But it is possible (and typical) to know of the singular causal relation without knowing the law or the relevant descriptions. Knowledge requires reasons, but these are available in the form of rough heteronomic generalizations, which are lawlike in that instances make it reasonable to expect other instances to follow suit without being lawlike in the sense of being indefinitely refinable. Applying these facts to knowledge of identities, we see that it is possible to know that a mental event is identical with some physical event without knowing which one (in the sense of being able to give it a unique physical description that brings it under a relevant law). Even if someone knew the entire physical history of the world, and every mental event were identical with a physical, it would not follow that he could predict or explain a single mental event (so described, of course).

Two features of mental events in their relation to the physical—causal dependence and nomological independence—combine, then, to dissolve what has often seemed a paradox, the efficacy of thought and purpose in the material world, and their freedom from law. When we portray events as perceivings, rememberings, decisions, and actions, we necessarily locate them amid physical happenings through the relation of cause and effect; but that same mode of portrayal insulates mental events, as long as we do not change the idiom, from the strict laws that can in principle be called upon to explain and predict physical phenomena.

Mental events as a class cannot be explained by physical science; particular mental events can when we know particular identities. But the explanations of mental events in which we are typically interested relate them to other mental events and conditions. We explain a man's free actions, for example, by appeal to his desires, habits, knowledge and perceptions. Such accounts of intentional behavior operate in a conceptual framework removed from the direct reach of physical law by describing both cause and effect, reason and action, as aspects of a portrait of a human agent. The anomalism of the mental is thus a necessary condition for viewing action as autonomous. I conclude with a second passage from Kant:

> It is an indispensable problem of speculative philosophy to show that its illusion respecting the contradiction rests on this, that we think of man in a different sense and relation when we call him free, and when we regard him as subject to the laws of nature. . . . It must therefore show that not only can both of these very well co-exist, but that both must be thought *as necessarily united* in the same subject. . . .[16]

NOTES

1. I was helped and influenced by Daniel Bennett, Sue Larson, and Richard Rorty, who are not responsible for the result. My research was supported by the National Science Foundation and the Center for Advanced Study in the Behavioral Sciences.

2. *Fundamental Principles of the Metaphysics of Morals*, T. K. Abbott, trans., (London: Longman, 1909), pp. 75–76.
3. These claims are defended in my "Actions, Reasons and Causes," *The Journal of Philosophy*, LX (1963): pp. 685–700 and in "Agency," in *Agent, Action, and Reason*, Robert Binkley, Richard Bronaugh, and Ausonio Marras, eds., (Oxford: Basil Blackwell, 1971), pp. 3–25.
4. In "Causal Relations," *The Journal of Philosophy* LXIV (1967): pp. 691–703, I elaborate on the view of causality assumed here. The stipulation that the laws be deterministic is stronger than required by the reasoning, and will be relaxed.
5. The point depends on assuming that mental events may intelligibly be said to have a location; but it is an assumption that must be true if an identity theory is, and here I am not trying to prove the theory but to formulate it.
6. I am indebted to Lee Bowie for emphasizing this difficulty.
7. Charles Taylor, "Mind-Body Identity, a Side Issue?," *The Philosophical Review* LXXVI (1967): p. 202.
8. J. J. C. Smart, "Sensations and Brain Processes," *The Philosophical Review* LXVIII (1959): pp. 141–56. The quoted passages are on pp. 163–65 of the reprinted version in *The Philosophy of Mind*, V. C. Chappell, ed., (Englewood Cliffs, NJ: Prentice-Hall,. Inc., 1962). For another example, see David K. Lewis, "An Argument for the Identity Theory," *The Journal of Philosophy* LXIII (1966): pp. 17–25. Here the assumption is made explicit when Lewis takes events as universals (p. 17, footnotes 1 and 2). I do not suggest that Smart and Lewis are confused, only that their way of stating the identity theory tends to obscure the distinction between particular events and kinds of events on which the formulation of my theory depends.
9. Jaegwon Kim, "On the Psycho-Physical Identity Theory," *American Philosophical Quarterly* III (1966): p. 231.
10. Ibid., pp. 227–28. Richard Brandt and Jaegwon Kim propose roughly the same criterion in "The Logic of the Identity Theory," *The Journal of Philosophy* LIV (1967): pp. 515–37. They remark that on their conception of event identity, the identity theory 'makes a stronger claim than merely that there is a pervasive phenomenal-physical correlation' (p. 518). I do not discuss the stronger claim.
11. Anomalous monism is more or less explicitly recognized as a possible position by Herbert Feigl, "The 'Mental' and the 'Physical,'" in *Concepts, Theories and the Mind–Body Problem*, vol. II, *Minnesota Studies in the Philosophy of Science* (Minneapolis: University of Minnesota Press, 1958); Sydney Shoemaker, "Ziff's Other Minds," *The Journal of Philosophy* LXII (1965): p. 589; David Randall Luce, "Mind-Body Identity and Psycho-Physical Correlation," *Philosophical Studies* XVII (1966): pp. 1–7; Charles Taylor, op. cit., p. 207. Something like my position is tentatively accepted by Thomas Nagel, "Physicalism," *The Philosophical Review* LXXIV (1965): pp. 339–56, and briefly endorsed by P. F. Strawson in *Freedom and the Will*, D. F. Pears, ed., (London: Macmillan, 1963), pp. 63–67.
12. The point that substitutivity of identity fails in the context of explanation is made in connection with the present subject by Norman Malcolm, "Scientific Materialism and the Identity Theory," *Dialogue* III (1964–65), pp. 123–24. See also my "Actions, Reasons and Causes," *The Journal of Philosophy* LX (1963): pp. 696–99; and "The Individuation of Events" in *Essays in Honor of Carl G. Hempel*, N. Rescher, et al., eds., (Dordrecht: D. Reidel, 1969).
13. The theme is developed in Roderick Chisholm, *Perceiving* (Ithaca, NY: Cornell University Press, 1957), chapter 11.
14. This view is accepted by Richard C. Jeffrey, "Goodman's Query," *The Journal of Philosophy* LXII (1966): p. 286 ff.; John R. Wallace, "Goodman, Logic, Induction," same journal and issue, p. 318; and John M. Vickers, "Characteristics of Projectible Predicates," *The Journal of Philosophy* LXIV (1967): p. 285. On pp. 328–29 and 286–87 of these journal issues respectively Goodman disputes the law-likeness of statements like 'All emerires are grue.' I cannot see, however, that he meets the point of my "Emeroses by Other Names," *The Journal of Philosophy* LXIII (1966): pp. 778–80.
15. The influence of W. V. Quine's doctrine of the indeterminacy of translation, as in chap. 2 of *Word and Object* (Cambridge, MA: MIT Press, 1960) is, I hope, obvious. In § 45 Quine develops the connection between translation and the propositional attitudes, and remarks that 'Brentano's thesis of the irreducibility of intentional idioms is of a piece with the thesis of indeterminacy of translation' (p. 221).
16. Op. cit., p. 76.

Special Sciences (or: The Disunity of Science as a Working Hypothesis)

Jerry A. Fodor

A typical thesis of positivistic philosophy of science is that all true theories in the special sciences should reduce to physical theories in the long run. This is intended to be an empirical thesis, and part of the evidence which supports it is provided by such scientific successes as the molecular theory of heat and the physical explanation of the chemical bond. But the philosophical popularity of the reductivist program cannot be explained by reference to these achievements alone. The development of science has witnessed the proliferation of specialized disciplines at least as often as it has witnessed their reduction to physics, so the widespread enthusiasm for reduction can hardly be a mere induction over its past successes.

I think that many philosophers who accept reductivism do so primarily because they wish to endorse the generality of physics vis-à-vis the special sciences: roughly, the view that all events which fall under the laws of any science are physical events and hence fall under the laws of physics.[1] For such philosophers, saying that physics is basic science and saying that theories in the special sciences must rede to physical theories have seemed to be two ways of saying the same thing, so that the latter doctrine has come to be a standard construal of the former.

In what follows, I shall argue that this is a considerable confusion. What has traditionally been called 'the unity of science' is a much stronger, and much less plausible, thesis than the generality of physics. If this is true it is important. Though reductionism is an empirical doctrine, it is intended to play a regulative role in scientific practice. Reducibility to physics is taken to be a *constraint* upon the acceptability of theories in the special sciences, with the curious consequence that the more the special sciences succeed, the more they ought to disappear. Methodological problems about psychology, in particular, arise in just this way: the assumption that the subject-matter of psychology is part of the subject-matter of physics is taken to imply that psychological theories must reduce to physical theories, and it is this latter principle that makes the trouble. I want to avoid the trouble by challenging the inference.

I

Reductivism is the view that all the special sciences reduce to physics. The sense of 'reduce to' is, however, proprietary. It can be characterized as follows.[2]

Let

(1) $S_1x \rightarrow S_2x$

be a law of the special science S. ((1) is intended to be read as something like 'all S_1 situations bring about S_2 situations.' I assume that a science is individuated largely by reference to its typical predicates, hence that if S is a special science 'S_1' and 'S_2' are not predicates of basic physics. I also assume that the 'all' which quantifies laws of the special sciences needs to be taken with a grain of salt; such laws are typically *not* exceptionless. This is a point to which I shall return at length.) A necessary and sufficient condition of the reduction of (1) to a law of physics is that the formulae (2) and (3) be laws, and a necessary and sufficient condition of the reduction of S to physics is that all its laws be so reducible.[3]

(2a) $S_1x \Delta P_1x$
(2b) $S_2x \Delta P_2x$
(3) $P_1x \rightarrow P_2x$

'P_1' and 'P_2' are supposed to be predicates of physics, and (3) is supposed to be a physical law. Formulae like (2) are often called 'bridge' laws. Their characteristic feature is that they contain predicates of both the reduced and the reducing science. Bridge laws like (2) are thus contrasted with 'proper' laws like (1) and (3).

From *Synthese* 28 (1974): pp. 97–115. Reprinted with permission of the author and of Kluwer Academic Publishers.

The upshot of the remarks so far is that the reduction of a science requires that any formula which appears as the antecedent or consequent of one of its proper laws must appear as the reduced formula in some bridge law or other.[4]

Several points about the connective '→' are in order. First, whatever other properties that connective may have, it is universally agreed that it must be transitive. This is important because it is usually assumed that the reduction of some of the special sciences proceeds via bridge laws which connect their predicates with those of intermediate reducing theories. Thus, psychology is presumed to reduce to physics via, say, neurology, biochemistry, and other local stops. The present point is that this makes no difference to the logic of the situation so long as the transitivity of '→' is assumed. Bridge laws which connect the predicates of *S* to those of *S** will satisfy the constraints upon the reduction of *S* to physics so long as there are other bridge laws which, directly or indirectly, connect the predicates of *S** to physical predicates.

There are, however, quite serious open questions about the interpretations of '→' in bridge laws. What turns on these questions is the respect in which reductivism is taken to be a physicalist thesis.

To begin with, if we read '→' as 'brings about' or 'causes' in proper laws, we will have to have some other connective for bridge laws, since bringing about and causing are presumably *a*symmetric, while bridge laws express symmetric relations. Moreover, if '→' in bridge laws is interpreted as any relation other than identity, the truth of reductivism will only guaranty the truth of a weak version of physicalism, and this would fail to express the underlying ontological bias of the reductivist program.

If bridge laws are not identity statements, then formulae like (2) claim at most that, by law, *x*'s satisfaction of a *P* predicate and *x*'s satisfaction of an *S* predicate are causally correlated. It follows from this that it is nomologically necessary that *S* and *P* predicates apply to the same things (i.e., that *S* predicates apply to a subset of the things that *P* predicates apply to). But, of course, this is compatible with a nonphysicalist ontology since it is compatible with the possibility that *x*'s satisfying *S* should not itself *be* a physical event. On this interpretation, the truth of reductivism does *not* guarantee the generality of physics vis-à-vis the special sciences since there are some events (satisfactions of *S* predicates) which fall in the domains of a special science (*S*) but not in the domain of physics. (One could imagine, for example, a doctrine according to which physical and psychological predicates are both held to apply to organisms, but where it is denied that the event which consists of an organism's satisfying a psychological predicate is, in any sense, a physical event. The up-shot would be a kind of psychophysical dualism of a non-Cartesian variety; a dualism of events and/or properties rather than substances.)

Given these sorts of considerations, many philosophers have held that bridge laws like (2) ought to be taken to express contingent event identities, so that one would read (2a) in some such fashion as 'every event which consists of *x*'s satisfying S_1 is identical to some event which consists of *x*'s satisfying P_1 and vice versa.' On this reading, the truth of reductivism would entail that every event that falls under any scientific law is a physical event, thereby simultaneously expressing the ontological bias of reductivism and guaranteeing the generality of physics vis-à-vis the special sciences.

If the bridge laws express event identities, and if every event that falls under the proper laws of a special science falls under a bridge law, we get the truth of a doctrine that I shall call 'token physicalism.' Token physicalism is simply the claim that all the events that the sciences talk about are physical events. There are three things to notice about token physicalism.

First, it is weaker than what is usually called 'materialism.' Materialism claims *both* that token physicalism is true *and* that every event falls under the laws of some science or other. One could therefore be a token physicalist without being a materialist, though I don't see why anyone would bother.

Second, token physicalism is weaker than what might be called 'type physicalism,' the doctrine, roughly, that every *property* mentioned in the laws of any science is a physical property. Token physicalism does not entail type physicalism because the contingent identity of a pair of events presumably does not guarantee the identity of the properties whose instantiation constitutes the events; not even where the event identity is nomologically necessary. On the other hand, if every event is the instantiation of a property, then type physicalism does entail token physicalism: two events will be identical when they consist of the instantiation of the same property by the same individual at the same time.

Third, token physicalism is weaker than reductivism. Since this point is, in a certain sense,

the burden of the argument to follow, I shan't labour it here. But, as a first approximation, reductivism is the conjunction of token physicalism with the assumption that there are natural kind predicate in an ideally completed physics which correspond to each natural kind predicates in any ideally completed special science. It will be one of my morals that the truth of reductivism cannot be inferred from the assumption that token physicalism is true. Reductivism is a sufficient, but not a necessary, condition for token physicalism.

In what follows, I shall assume a reading of reductivism which entails token physicalism. Bridge laws thus state nomologically necessary contingent event identities, and a reduction of psychology to neurology would entail that any event which consists of the instantiation of a psychological property is identical with some event which consists of the instantiation of some neurological property.

Where we have got to is this: reductivism entails the generality of physics in at least the sense that any event which falls within the universe of discourse of a special science will also fall within the universe of discourse of physics. Moreover, any prediction which follows from the laws of a special science and a statement of initial conditions will also follow from a theory which consists of physics and the bridge laws, together with the statement of initial conditions. Finally, since 'reduces to' is supposed to be an asymmetric relation, it will also turn out that physics is *the* basic science; that is, if reductivism is true, physics is the only science that is general in the sense just specified. I now want to argue that reductivism is too strong a constraint upon the unity of science, but that the relevantly weaker doctrine will preserve the desired consequences of reductivism: token physicalism, the generality of physics, and its basic position among the sciences.

II

Every science implies a taxonomy of the events in its universe of discourse. In particular, every science employs a descriptive vocabulary of theoretical and observation predicates such that events fall under the laws of the science by virtue of satisfying those predicates. Patently, not every true description of an event is a description in such a vocabulary. For example, there are a large number of events which consist of things having been transported to a distance of less than three miles from the Eiffel Tower. I take it, however, that there is no science which contains 'is transported to a distance of less than three miles from the Eiffel Tower' as part of its descriptive vocabulary. Equivalently, I take it that there is no natural law which applies to events in virtue of their being instantiations of the property *is transported to a distance of less than three miles from the Eiffel Tower* (though I suppose it is conceivable that there is some law that applies to events in virtue of their being instantiations of some distinct but co-extensive property). By way of abbreviating these facts, I shall say that the property *is transported . . .* does not determine a *natural kind*, and that predicates which express that property are not natural kind predicates.

If I knew what a law is, and if I believed that scientific theories consist just of bodies of laws, then I could say that P is a natural kind predicate relative to S if S contains proper laws of the form $P_x \rightarrow \alpha_x$ or $\alpha_x \rightarrow P_x$; roughly, the natural kind predicates of a science are the ones whose terms are the bound variables in its proper laws. I am inclined to say this even in my present state of ignorance, accepting the consequence that it makes the murky notion of a natural kind viciously dependent on the equally murky notions *law* and *theory*. There is no firm footing here. If we disagree about what is a natural kind, we will probably also disagree about what is a law, and for the same reasons. I don't know how to break out of this circle, but I think that there are interesting things to say about which circle we are in.

For example, we can now characterize the respect in which reductivism is too strong a construal of the doctrine of the unity of science. If reductivism is true, then *every* natural kind is, or is co-extensive with, a physical natural kind. (Every natural kind *is* a physical natural kind if bridge laws express property identities, and every natural kind is co-extensive with a physical natural kind if bridge laws express event identities.) This follows immediately from the reductivist premise that every predicate which appears as the antecedent or consequent of a law of the special sciences must appear as one of the reduced predicates in some bridge, together with the assumption that the natural kind predicates are the ones whose terms are the bound variables in proper laws. If, in short, some physical law is related to each law of a special science in the way that (3) is related to (1), then every natural kind predicate of a special science is related to a natural kind predicate

of physics in the way that (2) relates 'S_1' and 'S_2' to 'P_1' and 'P_2.'

I now want to suggest some reasons for believing that this consequence of reductivism is intolerable. These are not supposed to be knockdown reasons; they couldn't be, given that the question whether reductivism is too strong is finally an *empirical* question. (The world could turn out to be such that every natural kind corresponds to a physical natural kind, just as it could turn out to be such that the property *is transported to a distance of less than three miles from the Eiffel Tower* determines a natural kind in, say, hydrodynamics. It's just that, as things stand, it seems very unlikely that the world *will* turn out to be either of these ways.)

The reason it is unlikely that every natural kind corresponds to a physical natural kind is just that (a) interesting generalizations (e.g., counter-factual supporting generalizations) can often be made about events whose physical descriptions have nothing in common, (b) it is often the case that *whether* the physical descriptions of the events subsumed by these generalizations have anything in common is, in an obvious sense, entirely irrelevant to the truth of the generalizations, or to their interestingness, or to their degree of confirmation or, indeed, to any of their epistemologically important properties, and (c) the special sciences are very much in the business of making generalizations of this kind.

I take it that these remarks are obvious to the point of self-certification; they leap to the eye as soon as one makes the (apparently radical) move of taking the special sciences at all seriously. Suppose, for example, that Gresham's 'law' really is true. (If one doesn't like Gresham's law, then any true generalization of any conceivable future economics will probably do as well.) Gresham's law says something about what will happen in monetary exchanges under certain conditions. I am willing to believe that physics is general *in the sense that it implies that any event which consists of a monetary exchange* (hence any event which falls under Gresham's law) *has a true description in the vocabulary of physics and in virtue of which it falls under the laws of physics.* But banal considerations suggest that a description which covers all such events must be wildly disjunctive. Some monetary exchanges involve strings of wampum. Some involve dollar bills. And some involve signing one's name to a check. What are the chances that a disjunction of physical predicates which covers all these events (i.e., a disjunctive predicate which can form the right-hand side of a bridge law of the form 'x is a monetary exchange $\leftrightarrows$ . . .') expresses a physical natural kind? In particular, what are the chances that such a predicate forms the antecedent or consequent of some proper law of physics? The point is that monetary exchanges have interesting things in common; Gresham's law, if true, says what one of these interesting things is. But what is interesting about monetary exchanges is surely not their commonalities under *physical* description. A natural kind like a monetary exchange *could* turn out to be co-extensive with a physical natural kind; but if it did, that would be an accident on a cosmic scale.

In fact, the situation for reductivism is still worse than the discussion thus far suggests. For, reductivism claims not only that all natural kinds are co-extensive with physical natural kinds, but that the co-extensions are nomologically necessary: bridge laws are *laws.* So, if Gresham's law is true, it follows that there is a (bridge) law of nature such that 'x is a monetary exchange $\rightleftarrows$ x is P,' where P is a term for a physical natural kind. But, surely, there is no such law. If there were, then P would have to cover not only all the systems of monetary exchange that there *are*, but also all the systems of monetary exchange that there *could be;* a law must succeed with the counterfactuals. What physical predicate is a candidate for 'P' in 'x is a nomologically possible monetary exchange if P_x'?

To summarize: an immortal econophysicist might, when the whole show is over, find a predicate in physics that was, in brute fact, co-extensive with 'is a monetary exchange.' If physics is general—if the ontological biases of reductivism are true—then there must *be* such a predicate. But (a) to paraphrase a remark Donald Davidson made in a slightly different context, nothing but brute enumeration could convince us of this brute co-extensivity, and (b) there would seem to be no chance at all that the physical predicate employed in stating the co-extensivity is a natural kind term, and (c) there is still less chance that the co-extension would be lawful (i.e., that it would hold not only for the nomologically possible world that turned out to be real, but for any nomologically possible world at all).

I take it that the preceding discussion strongly suggests that economics is not reducible to physics in the proprietary sense of reduction involved in claims for the unity of science. There

is, I suspect, nothing special about economics in this respect; the reasons why economics is unlikely to reduce to physics are paralleled by those which suggest that psychology is unlikely to reduce to neurology.

If psychology is reducible to neurology, then for every psychological natural kind predicate there is a co-extensive neurological natural kind predicate, and the generalization which states this co-extension is a law. Clearly, many psychologists believe something of the sort. There are departments of 'psycho-biology' or 'psychology and brain science' in universities throughout the world whose very existence is an institutionalized gamble that such lawful co-extensions can be found. Yet, as has been frequently remarked in recent discussions of materialism, there are good grounds for hedging these bets. There are no firm data for any but the grossest correspondence between types of psychological states and types of neurological states, and it is entirely possible that the nervous system of higher organisms characteristically achieves a given psychological end by a wide variety of neurological means. If so, then the attempt to pair neurological structures with psychological functions is foredoomed. Physiological psychologists of the stature of Karl Lashley have held precisely this view.

The present point is that the reductivist program in psychology is, in any event, *not* to be defended on ontological grounds. Even if (token) psychological events are (token) neurological events, it does not follow that the natural kind predicates of psychology are co-extensive with the natural kind predicates of any other discipline (including physics). That is, the assumption that every psychological event is a physical event does not guaranty that physics (or, *a fortiori*, any other discipline more general than psychology) can provide an appropriate vocabulary for psychological theories. I emphasize this point because I am convinced that the make-or-break commitment of many physiological psychologists to the reductivist program stems precisely from having confused that program with (token) physicalism.

What I have been doubting is that there are neurological natural kinds co-extensive with psychological natural kinds. What seems increasingly clear is that, even if there is such a co-extension, it cannot be lawlike. For, it seems increasingly likely that there are nomologically possible systems other than organisms (namely, automata) which satisfy natural kind predicates in psychology, and which satisfy no neurological predicates at all. Now, as Putnam has emphasized, if there are any such systems, then there are probably vast numbers, since equivalent automata can be made out of practically anything. If this observation is correct, then there can be no serious hope that the class of automata whose psychology is effectively identical to that of some organism can be described by *physical* natural kind predicates (though, of course, if token physicalisms is true, that class can be picked out by some physical predicate or other). The upshot is that the classical formulation of the unity of science is at the mercy of progress in the field of computer simulation. This is, of course, simply to say that that formulation was too strong. The unity of science was intended to be an empirical hypothesis, defeasible by possible scientific findings. But no one had it in mind that it should be defeated by Newell, Shaw and Simon.

I have thus far argued that psychological reductivism (the doctrine that every psychological natural kind is, or is co-extensive with, a neurological natural kind) is not equivalent to, and cannot be inferred from, token physicalism (the doctrine that every psychological event is a neurological event). It may, however, be argued that one might as well take the doctrines to be equivalent since the only possible *evidence* one could have for token physicalism would also be evidence for reductivism: namely, the discovery of type-to-type psychophysical correlations.

A moment's consideration shows, however, that this argument is not well taken. If type-to-type psychophysical correlations would be evidence for token physicalism, so would correlations of other specifiable kinds.

We have type-to-type correlations where, for every n-tuple of events that are of the same psychological kind, there is a correlated n-tuple of events that are of the same neurological kind. Imagine a world in which such correlations are *not* forthcoming. What is found, instead, is that for every n-tuple of type identical psychological events, there is a spatiotemporally correlated n-tuple of type *distinct* neurological events. That is, every psychological event is paired with some neurological event or other, but psychological events of the same kind may be paired with neurological events of different kinds. My present point is that such pairings would provide as much support for token physicalism as type-to-type pairings do *so long as we are able to show that the type distinct neurological events paired with a given kind of psychological event are identical in respect*

of whatever properties are relevant to type-identification in psychology. Suppose, for purposes of explication, that psychological events are type identified by reference to their behavioral consequences.[5] Then what is required of all the neurological events paired with a class of type homogeneous psychological events is only that they be identical in respect of their behavioral consequences. To put it briefly, type identical events do not, of course, have *all* their properties in common, and type distinct events must nevertheless be identical in *some* of their properties. The empirical confirmation of token physicalism does not depend on showing that the neurological counterparts of type identical psychological events are themselves type identical. What needs to be shown is only that they are identical in respect of those properties which determine which kind of *psychological* event a given event is.

Could we have evidence that an otherwise heterogeneous set of neurological events have these kinds of properties in common? Of course we could. The neurological theory might itself explain why an *n*-tuple of neurologically type distinct events are identical in their behavioral consequences, or, indeed, in respect of any of indefinitely many other such relational properties. And, if the neurological theory failed to do so, some science more basic than neurology might succeed.

My point in all this is, once again, not that correlations between type homogeneous psychological states and type heterogeneous neurological states would prove that token physicalism is true. It is only that such correlations might give us as much reason to be token physicalists as type-to-type correlations would. If this is correct, then the epistemological arguments from token physicalism to reductivism must be wrong.

It seems to me (to put the point quite generally) that the classical construal of the unity of science has really misconstrued the *goal* of scientific reduction. The point of reduction is *not* primarily to find some natural kind predicate of physics co-extensive with each natural kind predicate of a reduced science. It is, rather, to explicate the physical mechanisms whereby events conform to the laws of the special sciences. I have been arguing that there is no logical or epistemological reason why success in the second of these projects should require success in the first, and that the two are likely to come apart *in fact* wherever the physical mechanisms whereby events conform to a law of the special sciences are heterogeneous.

III

I take it that the discussion thus far shows that reductivism is probably too strong a construal of the unity of science; on the one hand, it is incompatible with probable results in the special sciences, and, on the other, it is more than we need to assume if what we primarily want is just to be good token physicalists. In what follows, I shall try to sketch a liberalization of reductivism which seems to me to be just strong enough in these respects. I shall then give a couple of independent reasons for supposing that the revised doctrine may be the right one.

The problem all along has been that there is an open empirical possibility that what corresponds to the natural kind predicates of a reduced science may be a heterogeneous and unsystematic disjunction of predicates in the reducing science, and we do not want the unity of science to be prejudiced by this possibility. Suppose, then, that we allow that bridge statements may be of the form

$$(4) \quad Sx \leftrightarrows P_1x \vee P_2x \vee \ldots \vee P_nx,$$

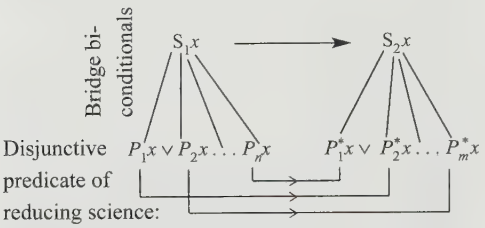
where '$P_1 \vee P_2 \vee \ldots \vee P_n$' is *not* a natural kind predicate in the reducing science. I take it that this is tantamount to allowing that at least some 'bridge laws' may, in fact, not turn out to be laws, since I take it that a necessary condition on a universal generalization being lawlike is that the predicates which constitute its antecedent and consequent should pick out natural kinds. I am thus supposing that it is enough, for purposes of the unity of science, that every law of the special sciences should be reducible to physics by bridge statements which express true empirical generalizations. Bearing in mind that bridge statements are to be construed as a species of identity statements, (4) will be read as something like 'every event which consists of x's satisfying S is identical with some event which consists of x's satisfying some or other predicate belonging to the disjunction '$P_1 \vee P_2 \vee \ldots \vee P_n$.'

Now, in cases of reduction where what corresponds to (2) is not a law, what corresponds to (3) will not be either, and for the same reason. Namely, the predicates appearing in the antecedent or consequent will, by hypothesis, not be natural kind predicates. Rather, what we will have is something that looks like (5) (see next page).

That is, the antecedent and consequent of the reduced law will each be connected with

a disjunction of predicates in the reducing science, and, if the reduced law is exceptionless, there will be laws of the reducing science which connect the satisfaction of each member of the disjunction associated with the antecedent to the satisfaction of some member of the disjunction associated with the consequent. That is, if $S_1x \rightarrow S_2x$ is

(5) Law of special science *X:*



exceptionless, then there must be some proper law of the reducing science which either states or entails that $P_1x \rightarrow P^*$ for some P^*, and similarly for P_2x through P_nx. Since there must be such laws, it follows that each disjunct of '$P_1 \vee P_2 \vee \ldots \vee P_n$' is a natural kind predicate, as is each disjunct of '$P^*_1 \vee P^*_2 \vee \ldots \vee P^*_n$.'

This, however, is where push comes to shove. For, it might be argued that if each disjunct of the P disjunction is lawfully connected to some disjunct of the P^* disjunction, it follows that (6) is itself a law.

(6) $P_1x \vee P_2x \vee \ldots \vee P_nx \rightarrow P^*_1x \vee P^*_2x \vee \ldots \vee P^*_nx$.

The point would be that (5) gives us $P^*_1x \rightarrow P^*_2x$, $P_2x \rightarrow P^*_mx$, etc., and the argument from a premise of the form $(P \supset R)$ and $(Q \supset S)$ to a conclusion of the form $(P \vee Q) \supset (R \vee S)$ is valid.

What I am inclined to say about this is that it just shows that 'it's a law that —' defines a non-truth functional context (or, equivalently for these purposes, that not all truth functions of natural kind predicates are themselves natural kind predicates). In particular, that one may not argue from 'it's a law that P brings about R' and 'it's a law that Q brings about S' to 'it's a law that P or Q brings about R or S.' (Though, of course, the argument from those premises to 'P or Q brings about R or S' *simpliciter* is fine.) I think, for example, that it is a law that the irradiation of green plants by sunlight causes carbohydrate synthesis, and I think that it is a law that friction causes heat, but I do not think that it is a law that (either the irradiation of green plants by sunlight or friction) causes (either carbohydrate synthesis or heat). Correspondingly, I doubt that 'is either carbohydrate synthesis or heat' is plausibly taken to be a natural kind predicate.

It is not strictly mandatory that one should agree with all this, but one denies it at a price. In particular, if one allows the full range of truth functional arguments inside the context 'it's a law that —,' then one gives up the possibility of identifying the natural kind predicates of a science with those predicates which appear as the antecedents or the consequents of its proper laws. (Thus (6) would be a proper law of physics which fails to satisfy that condition.) One thus inherits the need for an alternative construal of the notion of a natural kind, and I don't know what that alternative might be like.

The upshot seems to be this. If we do not require that bridge statements must be laws, then either some of the generalizations to which the laws of special sciences reduce are not themselves lawlike, or some laws are not formulable in terms of natural kinds. Whichever way one takes (5), the important point is that it is weaker than standard reductivism: it does not require correspondences between the natural kinds of the reduced and the reducing science. Yet it is physicalistic on the same assumption that makes standard reductivism physicalistic (namely, that the bridge statements express true token identities). But these are precisely the properties that we wanted a revised account of the unity of science to exhibit.

I now want to give two reasons for thinking that this construal of the unity of science is right. First, it allows us to see how the laws of the special sciences could reasonably have exceptions, and, second, it allows us to see why there are special sciences at all. These points in turn.

Consider, again, the model of reduction implicit in (2) and (3). I assume that the laws of basic science are strictly exceptionless, and I assume that it is common knowledge that the laws of the special sciences are not. But now we have a painful dilemma. Since '$\rightarrow$' expresses a relation (or relations) which must be transitive, (1) can have exceptions only if the bridge laws do. But if the bridge laws have exceptions, reductivism loses its ontological bite, since we can no longer say that every event which consists of the instantiation of an S predicate is identical with some event which consists of the instantiation of a P predicate. In short, given

the reductionist model, we cannot consistently assume that the bridge laws and the basic laws are exceptionless while assuming that the special laws are not. But we cannot accept the violation of the bridge laws unless we are willing to vitiate the ontological claim that is the main point of the reductivist program.

We can get out of this (*salve* the model) in one of two ways. We can give up the claim that the special laws have exceptions or we can give up the claim that the basic laws are exceptionless. I suggest that both alternatives are undesirable. The first because it flies in the face of fact. There is just no chance at all that the true, counter-factual supporting generalizations of, say, psychology, will turn out to hold in strictly each and every condition where their antecedents are satisfied. Even where the spirit is willing, the flesh is often weak. There are always going to be behavioral lapses which are physiologically explicable but which are uninteresting from the point of view of psychological theory. The second alternative is only slightly better. It may, after all, turn out that the laws of basic science have exceptions. But the question arises whether one wants the unity of science to depend upon the assumption that they do.

On the account summarized in (5), however, everything works out satisfactorily. A nomologically sufficient condition for an exception to $S_1x \rightarrow S_2x$ is that the bridge statements should identify some occurrence of the satisfaction of S_1 with an occurrence of the satisfaction of a P predicate which is not itself lawfully connected to the satisfaction of any P^* predicate. (I.e., suppose S_1 is connected to a P' such that there is no law which connects P' to any predicate which bridge statements associate with S_2. Then any instantiation of S_1 which is contingently identical to an instantiation of P' will be an event which constitutes an exception to $S_1x \rightarrow S_2x$.) Notice that, in this case, we need assume no exceptions to the laws of the *reducing* science since, by hypothesis, (6) *is not a law.*

In fact, strictly speaking, (6) has no status in the reduction at all. It is simply what one gets when one universally quantifies a formula whose antecedent is the physical disjunction corresponding to S_1 and whose consequent is the physical disjunction corresponding to S_2. As such, it will be true when $S_1 \rightarrow S_2$ is exceptionless and false otherwise. What does the work of expressing the physical mechanisms whereby n-tuples of events conform, or fail to conform, to $S_1 \rightarrow S_2$ is not (6) but the laws which severally relate elements of the disjunction $P_1 \vee P_2 \vee \ldots \vee P_n$ to elements of the disjunction $P^*_1 \vee P^*_2 \vee \ldots \vee P^*_n$. When there *is* a law which relates an event that satisfies one of the P disjuncts to an event which satisfies one of the P^* disjuncts, the pair of events so related conforms to $S_1 \rightarrow S_2$. When an event which satisfies a P predicate is *not* related by law to an event which satisfies a P^* predicate, that event will constitute an exception to $S_1 \rightarrow S_2$. The point is that none of the laws which effect these several connections need themselves have exceptions in order that $S_1 \rightarrow S_2$ should do so.

To put this discussion less technically: we could, if we liked, *require* the taxonomies of the special sciences to correspond to the taxonomy of physics by insisting upon distinctions between the natural kinds postulated by the former wherever they turn out to correspond to distinct natural kinds in the latter. This would *make* the laws of the special sciences exceptionless if the laws of basic science are. But it would also loose us precisely the generalizations which we want the special sciences to express. (If economics were to posit as many *kinds* of monetary systems as there are kinds of physical realizations of monetary systems, then the generalizations of economics *would* be exceptionless. But, presumably, only vacuously so, since there would be no generalizations left to state. Graham's law, for example, would have to be formulated as a vast, open disjunction about what happens in monetary system$_1$ or monetary system$_n$ under conditions which would themselves defy uniform characterization. We would not be able to say what happens in monetary systems *tourt court* since, by hypothesis, 'is a monetary system' corresponds to no natural kind predicate of physics.)

In fact, what we do is precisely the reverse. We allow the generalizations of the special sciences to *have* exceptions, thus preserving the natural kinds to which the generalizations apply. But since we know that the *physical* descriptions of the natural kinds may be quite heterogeneous, and since we know that the physical mechanisms which connect the satisfaction of the antecedents of such generalizations to the satisfaction of their consequents may be equally diverse, we expect both that there will be exceptions to the generalizations and that these exceptions will be 'explained away' at the level of the reducing science. This is one of the respects in which physics really is assumed to be bedrock science; exceptions to *its* generalizations (if there are any) had better be random, because there is nowhere 'further down' to go

in explaining the mechanism whereby the exceptions occur.

This brings us to why there are special sciences at all. Reductivism as we remarked at the outset, flies in the face of the facts about the scientific institution: the existence of a vast and interleaved conglomerate of special scientific disciplines which often appear to proceed with only the most token acknowledgment of the constraint that their theories must turn out to be physics 'in the long run.' I mean that the acceptance of this constraint, *in practice*, often plays little or no role in the validation of theories. Why is this so? Presumably, the reductivist answer must be *entirely* epistemological. If only physical particles weren't so small (if only brains were on the *out*side, where one can get a look at them), *then* we would do physics instead of paleontology (neurology instead of psychology; psychology instead of economics; and so on down). There is an epistemological reply; namely, that even if brains were out where they can be looked *at*, as things now stand, we wouldn't know what to look *for:* we lack the appropriate theoretical apparatus for the psychological taxonomy of neurological events.

If it turns out that the functional decomposition of the nervous system corresponds to its neurological (anatomical, biochemical, physical) decomposition, then there are only epistemological reasons for studying the former instead of the latter. But suppose there is no such correspondence? Suppose the functional organization of the nervous system cross-cuts its neurological organization (so that quite different neurological structures can subserve identical psychological functions across times or across organisms). Then the existence of psychology depends not on the fact that neurons are so sadly small, but rather on the fact that neurology does not posit the natural kinds that psychology requires.

I am suggesting, roughly, that there are special sciences not because of the nature of our epistemic relation to the world, but because of the way the world is put together: not all natural kinds (not all the classes of things and events about which there are important, counterfactual supporting generalizations to make) are, or correspond to, physical natural kinds. A way of stating the classical reductionist view is that things which belong to different physical kinds ipso facto can have no projectible descriptions in common; that if x and y differ in those descriptions by virtue of which they fall under the proper laws of physics, they must differ in those descriptions by virtue of which they fall under any laws at all. But why should we believe that this is so? Any pair of entities, however different their physical structure, must nevertheless converge in indefinitely many of their properties. Why should there not be, among those convergent properties, some whose lawful interrelations support the generalizations of the special sciences? Why, in short, should not the natural kind predicates of the special sciences *cross-classify* the physical natural kinds?[6]

Physics develops the taxonomy of its subject-matter which best suits its purposes: the formulation of exceptionless laws which are basic in the several senses discussed above. But this is not the only taxonomy which may be required if the purposes of science in general are to be served: e.g., if we are to state such true, counter-factual supporting generalizations as there are to state. So, there are special sciences, with their specialized taxonomies, in the business of stating some of these generalizations. If science is to be unified, then all such taxonomies must apply *to the same things.* If physics is to be basic science, then each of these things had better be a physical thing. But it is not further required that the taxonomies which the special sciences employ must themselves reduce to the taxonomy of physics. It is not required, and it is probably not true.

BIBLIOGRAPHY

Block, N. and Fodor, J., "What Psychological States Are Not," *Philosophical Review* 81 (1972): pp. 159–81.

Chomsky, N., *Aspects of the Theory of Syntax* (Cambridge, MA: MIT Press, 1965).

NOTES

I wish to express my gratitude to Ned Block for having read a version of this paper and for the very useful comments he made.

1. I shall usually assume that sciences are about events, in at least the sense that it is the occurrence of events that makes the laws of a science true. But I shall be pretty free with the relation between events, states, things and properties. I shall even permit myself some latitude in construing the relation between properties and predicates. I realize that all these relations are problems, but they aren't my problem in this paper. Explanation has to *start* somewhere, too.

2. The version of reductionism I shall be concerned with is a stronger one than many philosophers of science hold; a point worth emphasizing since my argument will be precisely that it is too strong to get away with. Still, I think that what I shall be attacking is what many people have in mind when they refer to the unity of science, and I suspect (though I shan't try to prove it) that many of the liberalized versions suffer from the same basic defect as what I take to be the classical form of the doctrine.
3. There is an implicit assumption that a science simply *is* a formulation of a set of laws. I think this assumption is implausible, but it is usually made when the unity of science is discussed, and it is neutral so far as the main argument of this paper is concerned.
4. I shall sometimes refer to 'the predicate which constitutes the antecedent or consequent of a law.' This is shorthand for 'the predicate such that the antecedent or consequent of a law consists of that predicate, together with its bound variables and the quantifiers which bind them.' (Truth functions of elementary predicates are, of course, themselves predicates in this usage.)
5. I don't think there is any chance at all that this is true. What is more likely is that type-identification for psychological states can be carried out in terms of the 'total states' of an abstract automaton which models the organism. For discussion, see Block and Fodor 1972.
6. As, by the way, the predicates of natural languages quite certainly do. For discussion, see Chomsky 1965.

Finding the Mind in the Natural World

Frank Jackson

Conceptual analysis played a prominent role in the defence of materialism mounted by the Australian materialists and their American ally David Lewis. It was how they found a place for the mind within the material world. The leading idea is encapsulated in the following argument schema:

1. Mental state M = occupant of functional role F. (By conceptual analysis)
2. Occupant of role F = brain state B. (By science)
3. Therefore, $M = B$. (By transitivity)

This schema gives the role of conceptual analysis in the Australian defence. But it does not tell us why conceptual analysis had to have a role in the defence. Indeed, the schema positively invites the thought that conceptual analysis was not needed. For to get the conclusion that $M = B$, all that is needed is the truth of the two premisses. It is not necessary that one of them be a conceptual truth. And I think, speaking more generally, that the Australian materialists left it unclear why materialists need to do some conceptual analysis. Nevertheless, I think that they were right that materialists need to do some conceptual analysis. This paper is a defence of this view. In a nutshell my argument will be that only by doing some conceptual analysis can materialists find a place for the mind in their naturalistic picture of the world. In a final section we will note the implications of our discussion for the knowledge argument.

In arguing for the necessity of conceptual analysis I am swimming against the tide. Current orthodoxy repudiates the role of conceptual analysis in the defence of materialism for at least three reasons. First, materialism is a doctrine in speculative metaphysics. And, runs the first reason, though conceptual analysis has a role in the philosophy of language and the study of concepts, it has no essential role when our subject is what the world is, at bottom, like. The second reason is that the history of conceptual analysis is the history of failure. For any proffered analysis someone clever always finds a counter-example. The final reason turns on the claim that we have learnt from Hilary Putnam and Saul Kripke about the necessary a posteriori, and that tells us that there can be necessary connections that, precisely by virtue of being a posteriori, are not revealed by or answerable to conceptual analysis. The materialist should, according to this line of thought, hold that the connection between the mental and the material or physical is a necessary a posteriori one, and so not a matter accessible via conceptual

From R. Casati, B. Smith, and G. White, eds., *Philosophy and the Cognitive Sciences* (Holder-Pichler-Tempsky, 1994). Reprinted with permission of the author.

analysis. During the course of the discussion we will see how to reply to each of these objections to the need for conceptual analysis in the defence of materialism.

The first step in our defence of the materialists' need for conceptual analysis is to note that materialism is a piece of what I will call serious metaphysics, and that, like any piece of serious metaphysics, it faces the location problem.

1. The Location Problem

Metaphysics is about what there is and what it is like. But it is concerned not with any old shopping list of what there is and what it is like. Metaphysicians seek a comprehensive account of some subject matter—the mind, the semantic, or, most ambitiously, everything—in terms of a limited number of more or less fundamental notions. Some who discuss the debate in the philosophy of mind between dualism and monism complain that *each* position is equally absurd. We should be *pluralists*. Of course we should be pluralists in some sense or other. However, if the thought is that any attempt to explain it all, or to explain it all as far as the mind is concerned, in terms of some limited set of fundamental ingredients is mistaken in principle, then it seems to me that we are being, in effect, invited to abandon serious metaphysics in favour of drawing up big lists. And we know we can do better than that. At least some of the diversity in our world conceals an underlying identity of ingredients. The diversity is a matter of the same elements differently selected and arranged. But if metaphysics seeks comprehension in terms of limited ingredients, it is continually going to be faced with the problem of location. Because the ingredients *are* limited, some putative features of the world are not going to appear explicitly in the story told in the favoured terms. The question then will be whether the features nevertheless figure *implicitly* in the story. Serious metaphysics is simultaneously discriminatory and putatively complete, and the combination of these two facts means that there is bound to be a whole range of putative features of our world up for either elimination or location.

What then is it for some putative feature to have a place in the story some metaphysic tells in its favoured terms? One answer is for the feature to be entailed by the story told in the favoured terms. Perhaps the story includes information about mass and volume in so many words, but nowhere mentions density by name. No matter—density facts are entailed by mass and volume facts. Or perhaps the story in the favoured terms says that many of the objects around us are nothing but aggregations of molecules held in a lattice-like array by various intermolecular forces. Nowhere in the story in the favoured terms is there any mention of solidity. Should we then infer that nothing is solid, or at any rate that this particular metaphysic is committed to nothing being solid? Obviously not. The story in the favoured terms will, we may suppose, tell us that these lattice-like arrays of molecules exclude each other, the inter-molecular forces being such as to prevent the lattices encroaching on each other's spaces. And *that* is what we understand by solidity. That's what it takes, according to our concept, to be solid. Or at least it is near enough. Perhaps pre-scientifically we might have been tempted to insist that being solid required being everywhere dense in addition to resisting encroachment. But resisting encroachment explains the stubbing of toes quite well enough for it to be pedantic to insist on anything more in order to be solid. Hence, solidity gets a location or place in the molecular story about our world by being entailed by that story, and we see this by asking ourselves about our concept of solidity in the sense of asking what it takes to be solid.

Thus, one way materialists can show that the psychological has a place in their world view is by showing that the psychological story is entailed by the story about the world told in the materialists' favoured terms. We will see, however, that it is not just one way; it is the one and only way.

2. Completeness and Supervenience

Materialism is the very opposite of a 'big list' metaphysics. It is highly discriminatory, operating in terms of a small set of favoured particulars, properties and relations, typically dubbed 'physical'—hence its other name, 'physicalism'; and it claims that a complete story, or anyway a complete story of everything contingent, including everything psychological, about our world can in principle be told in terms of these physical particulars, properties and relations alone. Only then is materialism interestingly different from dual attribute theories of mind.

Now what, precisely, is a complete story? We can make a start by noting that one particularly

clear way of showing *incompleteness* is by appeal to independent variation. What shows that three co-ordinates do not provide a complete account of location in space-time is that we can vary position in space-time while keeping any three co-ordinates constant. Hence, an obvious way to approach completeness is in terms of the lack of independent variation. But, of course, lack of independent variation is supervenience: position in space-time supervenes on the four co-ordinates. So the place to look when looking for illumination regarding the sense in which materialism claims to be complete, and, in particular, to be complete with respect to the psychological, is at various supervenience theses.[1]

Now materialism is not just a claim about the completeness of the physical story concerning certain individuals or particulars in our world. It claims completeness concerning the world itself, concerning, that is, the total way things are. Accordingly, we need to think of the supervenience base as consisting of possible worlds—complete ways things might be. We need, accordingly, to look to global supervenience theses, an example of which is

(I) Any two possible worlds that are physical duplicates (physical property, particular and relation for physical property, particular and relation identical) are duplicates *simpliciter*.

But (I) does not capture what the materialists have in mind. Materialism is a claim about our world, the actual world, to the effect that its physical nature exhausts all its nature, whereas (I) is a claim about worlds in general. A more restricted supervenience thesis in which our world is explicitly mentioned is:

(II) Any world that is a physical duplicate of our world is a duplicate *simpliciter* of our world.

However, materialists can surely grant that there is a possible world physically exactly like ours but which contains as an addition a lot of mental life sustained in non-physical stuff, as long as they insist that this world is not our world. Consider the view of those theists that hold that materialism is the correct account of earthly existence but it leaves out of account the after-life. When we die our purely material psychology is reinstated in purely non-physical stuff. Surely materialists can grant that these theists are right about some world, some way things might be, as long as they insist that it is *not* our world, not the way things actually are. Hence, materialists are not committed to (II).

The trouble with (II) is that it represents materialists' claims as more wide ranging than they in fact are. What we need is something like (II) but that limits itself to worlds more nearly like ours, or at least more nearly like ours on the materialists' conception of what our world is like. I suggest.

(III) Any world that is a *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world.

What is a minimal physical duplicate? Think of a recipe for making scones. It tells you what to do, but not what *not* to do. It tells you to add butter to the flour but does not tell you not to add whole peppercorns to the flour. Why doesn't it? Part of the reason is that no one would think to add them unless explicitly told to. But part of the reason is logical. It is impossible to list all the things *not* to do. There are indefinitely many of them. Of necessity the writers of recipes rely on an intuitive understanding of an implicitly included 'stop' clause in their recipes. A minimal physical duplicate of our world is what you would get if you used the physical nature of our world (including of course its physical laws) as a recipe in this sense for making a world.

We arrived at (III) by eliminating alternatives. But we can give a positive argument for the conclusion that the materialist is committed to (III). Suppose that (III) is false; then there is a difference in nature between our world and some minimal physical duplicate of it. But then either our world contains some nature that the minimal physical duplicate does not, or the minimal physical duplicate contains some nature that our world does not. The second is impossible because the extra nature would have to be non-physical (as our world and the duplicate are physically identical), and the minimal physical duplicate contains no non-physical nature by definition. But if our world contains some nature that the duplicate does not, this nature must be non-physical (as our world and the duplicate are physically identical). But then materialism would be false, for our world would contain some non-physical nature. Hence, if (III) is false, materialism is false—that is to say, materialism is committed to (III).

3. From (III) to Entry by Entailment

Given that (III) follows from materialism, there is a straightforward and familiar argument to show that if materialism is true, then the psychological story about our world is entailed by the physical story about our world.

We can think of a statement as telling a story about the way the world is, and as being true inasmuch as the world is the way the story says it is. Let ϕ be the statement which tells the rich, complex and detailed physical story that is true at the actual world and all and only the minimal physical duplicates of the actual world, and false elsewhere. Let Π be any true statement entirely about the psychological nature of our world: Π is true at our world, and every world at which Π is false differs in some psychological way from our world. If (III) is true, every world at which ϕ is true is a duplicate *simpliciter* of our world, and so a fortiori a psychological duplicate of our world. But then every world at which Π is true is a world at which Π is true—that is, ϕ entails Π.

We have thus derived what we might call the *entry by entailment thesis:* a putative psychological fact has a place in the materialists' world view if and only if it is entailed by the physical story about the world. The one and only way of getting a place is by entailment.

4. From Entry by Entailment to Conceptual Analysis

How does entry by entailment show the importance of conceptual analysis? If ϕ entails Π, what makes ϕ true also makes Π true (at least when ϕ and Π are contingent). But what makes ϕ true is the physical way our world is. Hence, the materialist is committed to each and every psychological statement being made true by a purely physical way our world is. But it is the very business of conceptual analysis to address which matters framed in terms of one set of terms and concepts are made true by which matters framed in a different set of terms and concepts. For instance, when we seek an analysis of knowledge in terms of truth, belief, justification, causation and so on, we seek an account of how matters described in terms of the latter notions make true matters described in terms of the former. When we seek an account of reference, we seek an account of the kinds of causal and descriptive facts which make it true that a term names an object. When and if we succeed, we will have an account of what makes it true that 'Moses' names Moses in terms of, among other things, causal links between uses of the word and Moses himself. And so on and so forth.

How could the a priori reflections on, and intuitions about, possible cases so distinctive of conceptual analysis be relevant to, for instance, the causal theory of reference? Well, the causal theory of reference is a theory about the conditions under which, say, 'Moses' refers to a certain person. But that is nothing other than a theory about the possible situations in which 'Moses' refers to that person, and the possible situations in which 'Moses' does not refer to that person. Hence, intuitions about various possible situations—the meat and potatoes of conceptual analysis—are bound to hold centre stage. (This is particularly true when the test situations cannot be realised. We cannot, for instance, make twin earth to check empirically what we would say about whether XYZ is water.)

The alternative is to *invent* our answers. Faced with the question, say, of whether the physical way things are makes true the belief way things are, we *could* stipulate the conditions under which something counts as a belief in such a way as to ensure that there are beliefs, or, if we preferred, that there are no beliefs. But that would not bear on whether beliefs according to *our* concept have a place in the materialists' picture of things, only on whether beliefs according to the stipulated concept have a place. In order to address the question of whether beliefs as we understand them have a place, what else can we do but consult and be guided by our honed intuitions about what counts as a belief? Would it be better to invent, or to go by what seems counter-intuitive?

I should emphasise, though, that a sensible use of conceptual analysis will allow a limited but significant place for a posteriori stipulation. We mentioned earlier the example of finding a place for solidity in the molecular picture of our world, and the fact that what the molecular picture vindicates is the existence of solid bodies according to a conception of solidity cashed out in terms of mutual exclusion rather than in terms of the conjunction of mutual exclusion and being everywhere dense. For our day to day traffic with objects, it is the mutual exclusion that matters, and accordingly it is entirely reasonable to rule that mutual exclusion is enough for solidity. The role of conceptual analysis of

K-hood is not always to settle on a nice, neat, *totally* a priori list of necessary and sufficient conditions for being a *K*—indeed, that is the task that has so often been beyond us. It is rather to guide us in dividing up the cases that clearly are not cases of a *K*, from the cases that a principle of charity might lead us to allow as cases of a *K*. Then, armed with this information, we are in a position to address the question of whether some inventory of fundamental ingredients does, or does not, have a place for *K*s.

I should also emphasise that the contention is not that a priori reflection on possible cases gives us new information, let alone some sort of infallible new information, about what the world is like. The reflection is a priori in the sense that we are not consulting our intuitions about what would *happen* in certain possible cases—it is not like the famous thought experiments in science—rather we are consulting our intuitions about how to *describe* certain possible cases. And what we learn (in the sense of making explicit) is not something new about what the world is like, but something about how, given what the world is like as described in one set of terms, it should be described in some other set of terms. Perhaps the point is clearest in the example about finding solidity in the molecular account of the objects around us. Reflection on our concept of solidity tells us that the molecular account includes solidity, but it does not tell us that solidity is an addition to what appears in the molecular account of objects, let alone an infallible one.

5. The Objection from the Necessary a Posteriori

It might well be urged that the argument given above from (III) to the conclusion that ϕ entails Π is undermined by the existence of necessary a posteriori truths. The objection can be put in two different ways. Consider

> Over 60% of the Earth is covered by H_2O.
>
> Therefore, over 60% of the Earth is covered by water.

One way of putting the objection is that although every world where the premise is true is a world where the conclusion is true, the argument is not valid because the premise does not entail the conclusion in the relevant sense. It is not possible to move a priori from the premise to the conclusion. The premise fixes the conclusion without entailing it, as it is sometimes put. Likewise, for all we have shown by the considerations based on (III), ϕ fixes Π but does not entail it.

This way of putting the objection makes it sound like a quarrel over terminology. It invites the response of distinguishing entailment *simpliciter*, the notion cashed out simply in terms of being necessarily truth-preserving, from a priori or, as it is sometimes called, conceptual, entailment, the latter being the notion tied to a priori deducibility. But the real objection, of course, is that the necessarily truth-preserving nature of the passage from 'Over 60% of the Earth is covered by H_2O' to 'Over 60% of the Earth is covered by water' is not one that can in principle be revealed by conceptual analysis. Reflection on, and intuitions about, possible cases and concepts, unless supplemented by the a posteriori information that water is H_2O, will get you nowhere. Materialists, it seems, can allow that (III) forces them to admit a necessarily truth-preserving passage from to ϕ to Π, without allowing a role for conceptual analysis. They can simply insist that the entailment from ϕ to Π is an a posteriori one.

We will see, however, that acknowledging the necessary a posteriori does not alter matters in any essential respects as far as the importance of conceptual analysis goes. The argument to this conclusion turns on a negative claim about the nature of the necessity possessed by the necessary a posteriori, and a consequent view about the role of conceptual analysis, in the sense of intuitions about possibilities, in the detection of the necessary a posteriori.

6. The Necessity of the Necessary A Posteriori

There are two different ways of looking at the distinction between necessary a posteriori statements like 'Water = H_2O' and necessary a priori ones like '$H_2O = H_2O$' (all necessary modulo worlds where there is no water, of course). You might say that the latter are analytically or conceptually or logically (in some wide sense not tied to provability in a formal system) necessary, whereas the former are metaphysically necessary, meaning by the terminology that we are dealing with two senses of 'necessary' in somewhat the way that we are when we contrast logical necessity with nomic necessity. On this approach, the reason the necessity of water's being H_2O is not available a priori is that its necessity is not the kind that is available a priori.

I think, as against this view, that it is a mistake to hold that the necessity possessed by 'Water = H_2O' and 'If over 60% of the Earth is covered by H_2O, then over 60% of the Earth is covered by water' is different from that possessed by 'Water = water' and 'If over 60% of the Earth is covered by H_2O, then over 60% of the Earth is covered by H_2O.' Just as Quine insists that numbers and tables exist in the very same sense, I think that we should insist that water's being H_2O and water's being water are necessary in the very same sense.

My reason for holding that there is one sense of necessity here relates to what it was that convinced us that 'Water = H_2O' is necessarily true. What convinced us were the arguments of Saul Kripke and Hilary Putnam about how to *describe* certain possibilities, rather than arguments about what is possible per se. Kripke and Putnam convinced us that a world where XYZ plays the water role—that is, satisfies enough of (but how much is enough is vague): filling the oceans, being necessary for life, being colourless, being called 'water' by experts, being of a kind with the exemplars we are acquainted with, and so on—did not warrant the description 'world where water is XYZ,' and the stuff correctly described as water in a counterfactual world is the stuff—H_2O—which fills the water role in the actual world. The key point is that the right way to describe a counterfactual world sometimes depends in part on how the actual world is, and not solely on how the counterfactual world is in itself. The point is not one about the space of possible worlds in some newly recognised sense of 'possible,' but instead one about the role of the actual possible world in determining the correct way to describe certain counterfactual possible worlds—in the sense of 'possible' already recognised.

All this was, it seems to me, an exercise in conceptual analysis. We had an old theory about the meaning of 'water,' namely, that it meant 'that which fills the water role,' a theory that was refuted by appealing to our intuitions about how to describe possible worlds in which something different from that which actually fills the water role fills the water role. We became convinced of a new theory—again by reflection on possible cases, the meat and potatoes of conceptual analysis—according to which 'water' is a rigid designator of the stuff that fills the water role in the actual world. At no time did we have to recognise a new sort of possibility, only a new way for something in some counterfactual situation to count as a *K*, namely, by virtue not solely of how things are in that counterfactual situation, but in part in virtue of how things actually are.

If this is right, the inference

> Over 60% of the Earth is covered by H_2O.
>
> Therefore, over 60% of the Earth is covered by water.

is not an example of an a posteriori entailment that shows the irrelevance of conceptual analysis to the question of whether an a posteriori entailment holds. For it is conceptual analysis that tells us, in light of the fact that H_2O fills the water role, that the entailment holds.

7. Two-Dimensionalism and the Knowledge Argument

I have argued that materialists must hold that the complete story about the physical nature of our world given by ϕ entails everything about our psychology, and that such a position cannot be maintained independently of the results of conceptual analysis. But it is quite another question whether they must hold that ϕ a priori entails everything about our psychology, including its phenomenal side, and so quite another question whether they must hold that it is in principle possible to deduce from the full physical story alone what it is like to see red or smell a rose—the key assumption in the knowledge argument that materialism leaves out qualia. I will conclude by noting how the two-dimensional treatment of the necessary a posteriori—the obvious treatment of the necessary a posteriori for anyone sympathetic to the view that such necessity is not a new sort of necessity—means that materialists are committed to the a priori deducibility of the phenomenal from the physical.

If the explanation of the a posteriori nature of the necessary a posteriori does not lie in the special necessity possessed, where does it lie? Two dimensionalists insist that the issue is an issue about sentences, and not about propositions, or at least not propositions thought of as sets of possible worlds. For, by the conclusion that we are not dealing with a new sort of necessity, the set of worlds where water is water is the very same set as the set where water is H_2O, and so, by Leibnitz's Law, there is no question of the proposition that water is water differing from the proposition that water is H_2O in that one is, and one is not, necessary a posteriori. Their contention is that there are

sentences such that the proposition expressed by them depends on the context of utterance.[2] We understand them in that we know how the proposition expressed depends on the context, but if we do not know the relevant fact about the context, we will not know the proposition expressed. (In Robert Stalnaker's terminology, we know the propositional concept but not the proposition; in David Kaplan's, we know the character but not the content.[3]) Consider 'Over 60% of the Earth is covered by water.' Because 'water' is a rigid designator whose reference is fixed by 'the stuff that fills the water role,' someone who does not know what that stuff is does not know which proposition the sentence expresses, but they understand the sentence by virtue of knowing how the proposition expressed depends on how things actually are, and, in particular, this being the relevant contextual matter in this case, on what actually fills the water role. The explanation of the necessary a posteriori status of 'If over 60% of the Earth is covered by H_2O, then over 60% of the Earth is covered by water' then runs as follows. The proposition expressed by the sentence 'Over 60% of the Earth is covered by H_2O,' is the same as the proposition expressed by 'Over 60% of the Earth is covered by water,' and so the proposition expressed by the conditional sentence is a priori and necessary. But consistent with what is required to count as understanding the conditional sentence, it is contingent and a posteriori that it expresses a necessary a priori proposition.

I should emphasise that this does not mean that people who fully understand a sentence like 'Over 60% of the Earth is covered by water' but do not know that water is H_2O do not, in some perfectly natural sense, know the conditions under which what they are saying is true.[4] True, full understanding of the sentence does not in itself yield which proposition is expressed by the sentence, but knowledge of the way in which the proposition expressed depends on context, combined with knowledge of the truth conditions of the various propositions, does enable them to say when the sentence they produce is true. For their knowledge about how the proposition expressed depends on context together with the conditions under which the various propositions are true is given in the following array:

> If H_2O fills the water role, then 'Over 60% of the Earth is covered by water' expresses a proposition that is true if over 60% of the Earth is covered by H_2O.
>
> If XYZ fills the water role, then 'Over 60% of the Earth is covered by water' expresses a proposition that is true if over 60% of the Earth is covered by XYZ.
>
> If—fills the water role, then 'Over 60% of the Earth is covered by water' expresses a proposition that is true if over 60% of the Earth is covered by—.

For each distinct, context-giving, antecedent, a distinct proposition is expressed by the sentence. Nevertheless, simple inspection of the array shows that the sentence is true if over 60% of the Earth is covered by the stuff that fills the water role. That is the sense in which the fully understanding producer of the sentence knows when the sentence is true.[5]

Now, to return to the main plot, although understanding alone does not necessarily give the proposition expressed by certain sentences—that is how they can be necessary and yet this fact be in principle not accessible to understanding plus acumen alone, that is how they can be necessary a posteriori—understanding alone does give us the way the proposition expressed depends on context; and that fact is enough for us to move a priori from, for example, sentences about the distribution of H_2O combined with the right context-giving sentences, to information about the distribution of water. Consider, for instance, a supplementation of our earlier inference:

(1) Over 60% of the Earth is covered by H_2O.
(2) H_2O fills the water role.
(3) Therefore, over 60% of the Earth is covered by water.

Although, as noted earlier, the passage from (1) to (3) is necessarily truth-preserving but a posteriori being an a posteriori entailment, the passage from (1) and (2), to (3) is a priori. And it is so because, although our understanding of 'Over 60% of the Earth is covered by H_2O' does not in itself yield the proposition expressed by the sentence, it yields how the proposition depends on context, and (2) gives that context. (2) gives the relevant fact about how things are 'outside the head.' We did not know that (1) entailed (3) until we learnt (2), because we did not, and could not, have known that (1) and (3) express the same proposition until we learnt (2). But as soon as we learn (2), we have the wherewithal, if we are smart enough, to move a priori to (3).

The point, then, is that the necessary a posteriori nature of 'Water = H_2O' does not mean

that the fact that the H_2O way things are entails the water way things are is not answerable to our grasp of the relevant concepts plus acumen. It means, rather, that we need to tell a rich enough story about the H_2O way things are, a story that includes the crucial contextual information, before we can move from the H_2O way things are to the water way they are using our grasp of the concepts alone.

More generally, the two-dimensional way of looking at the necessary a posteriori means that even if the entailment the materialist is committed to from some physical story about the world to the full psychological story is a posteriori, there is still an a priori story tellable about how the story in physical terms about our world makes true the story in psychological terms about our world. Although understanding may not, even in principle, be enough to yield the proposition expressed by the physical story, understanding and logical acumen is enough to yield how the proposition expressed depends on context. But, of course, the context is, according to the materialist, entirely physical. The context concerns various matters about the nature of the actual world, and that nature is capturable in entirely physical terms according to the materialist. Hence, the materialist is committed to there being an a priori story to tell about how the physical way things are makes true the psychological way things are. But the story may come in two parts. It may be that one part of the story says which physical way things are, ϕ_1, makes some psychological statement true, and the other part of the story, the part that tells the context, says which different physical way things are, ϕ_2, makes it the case that it is ϕ_1 that makes the psychological statement true. What will be a priori accessible is that ϕ_1 and ϕ_2 together make the psychological statement true.[6]

REFERENCES

Horgan, Terence, "Supervenience and Microphysics," *Pacific Philosophical Quarterly* 63 (1982): pp. 29–43.

Kaplan, David, "Dthat," in *Syntax and Semantics*, vol. 9, P. Cole, ed., (New York: Academic Press, 1978).

Lewis, David, "New Work for a Theory of Universals," *Australasian Journal of Philosophy* 61 (1983): pp. 343–377.

Stalnaker, Robert C., "Assertion," in *Syntax and Semantics*, vol. 9, P. Cole, ed., (New York: Academic Press, 1978), pp. 315–32.

NOTES

1. What follows is one version of a familiar story. See, for example, T. Horgan 1982 and D. Lewis 1983.
2. I take it that what follows is a sketch of the approach suggested by the version of two-dimensionalism in Stalnaker 1978.
3. Stalnaker 1978 and Kaplan 1978.
4. I am indebted here to David Lewis and David Chalmers.
5. This observation bears on the dispute about whether Earthians and Twin Earthians believe alike. Although the sentence 'Water is plentiful' expresses different propositions in the mouths of the Earthians and the Twin Earthians, they agree about when the sentence is true, and so in *that* sense agree in belief.
6. I am indebted to Lloyd Humberstone, David Chalmers, David Lewis, Michael Smith, and Philip Pettit.

The Many Problems of Mental Causation

Jaegwon Kim

Giving an account of mental causation—in particular, explaining how it is possible for the mental to exercise causal influences in the physical world—has been one of the main preoccupations in the philosophy of mind over the past two decades. The problem of course is not new: as we learn early in our philosophy classes, Descartes was confronted forcefully by his contemporaries on this issue,[1] to explain how there could be causal transactions between minds and bodies. But this does not mean that Descartes' problem is our problem. His problem, as his contemporaries saw, was to show just how his all-too-commonsensical thesis of mind–body interaction was tenable within his ontology of two radically diverse domains of substances, minds and bodies. In his replies, Descartes hemmed and hawed, and was ultimately unable to produce an effective response. Many of his contemporaries, like Leibniz and Malebranche, chose to abandon mental causation in favor of substantival dualism. In staying with mental causation to the end, however, Descartes showed a healthy and commendable respect for philosophical commonsense—more so than many of his major philosophical rivals who opted for radical and implausible solutions—and I believe we should remember him for this as well as for his much publicized failure to reconcile mental causation with his ontology. In any case substance dualism is not the source of our current worries about mental causation; substantival minds are no longer a live philosophical option for most of us.

Philosophical problems do not arise in a vacuum. Typically they emerge when we come to see a conflict among the assumptions and presumptions that we explicitly or tacitly accept, or commitments that command our presumptive respect. The seriousness of a philosophical problem therefore depends on two related questions: First, how deep is our attachment to the assumptions and commitments that give rise to the apparent conflict? Second, how easy or difficult is it to bring the conflicting assumptions into an acceptable reconciliation? The process of reconciliation may require serious modifications to our original commitments. Short of abandoning the entire framework of the existing commitments, compromises must be negotiated. There are no free lunches in philosophy any more than in real life.

In this lecture I want to set out, in what to my mind is the simplest and starkest way, how our principal current problem of mental causation arises. In saying this, I do not want to imply that there is a single problem of mental causation. In fact, as we will shortly see, several different sets of assumptions and principles that many of us find plausible can make trouble for mental causation. I will first describe three sources that seem to generate difficulties for mental causation. This means that we are faced with at least three distinct problems of mental causation. However, in the rest of this lecture, I will focus on one particular version of the third of these problems ('the exclusion problem'). This problem arises from what I will call 'the supervenience argument.' This, I claim, is our principal problem of mental causation. In referring to this as 'our' problem of mental causation, what I mean to suggest is that it is a problem that arises for anyone with the kind of broadly physicalist outlook that many philosophers, including myself, find compelling or, at least, plausible and attractive. In contrast, the other two problems (the mental anomaly problem and the extrinsicness problem) are not essentially tied to physicalism. They are largely independent of physicalist commitments and can arise outside the physicalist framework. As we will see, the exclusion problem is distinctive in that it strikes

From *Mind in a Physical World* (MIT Press, 1998), pp. 29–47.

at the very heart of physicalism, and I believe that the supervenience argument captures the essence of the difficulties involved. The fundamental problem of mental causation for us, then, is to answer this question: How is it possible for the mind to exercise its causal powers in a world that is fundamentally physical?

Let me begin with some reasons for wanting to save mental causation—why it is important to us that mental causation is real (some will say that its existence is an ultimate, nonnegotiable commitment). First, the possibility of human agency evidently requires that our mental states—our beliefs, desires, and intentions—have causal effects in the physical world: in voluntary actions our beliefs and desires, or intentions and decisions, must somehow cause our limbs to move in appropriate ways, thereby causing the objects around us to be rearranged. That is how we manage to cope with our surroundings, write philosophy papers, build bridges and cities, and make holes in the ozone layers. Second, the possibility of human knowledge presupposes the reality of mental causation: perception, our sole window on the world, requires the causation of perceptual experiences and beliefs by physical objects and events around us. Reasoning, by which we acquire new knowledge and belief from the existing fund of what we already know or believe, involves the causation of new belief by old belief; more generally, causation arguably is essential to the transmission of evidential groundedness. Memory is a complex causal process involving interactions between experiences, their physical storage, and retrieval in the form of belief. If you take away perception, memory, and reasoning, you pretty much take away all of human knowledge. To move on, it seems plain that the possibility of psychology as a theoretical science capable of generating law-based explanations of human behavior depends on the reality of mental causation: mental phenomena must be capable of functioning as indispensable links in causal chains leading to physical behavior. A science that invokes mental phenomena in its explanations is presumptively committed to their causal efficacy; for any phenomenon to have an explanatory role, its presence or absence in a given situation must make a difference—a *causal difference.*

It is no wonder then that for most philosophers the causal efficacy of the mental is something that absolutely cannot be given away no matter how great the pressures are from other quarters. Jerry Fodor is among these philosophers; he writes:

> . . . if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying . . . , if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.[2]

If mental causation is only an illusion, that perhaps is not the end of the world, but it surely seems like the end of a world that includes Fodor and the rest of us as agents and cognizers. The problem of determinism threatens human agency, and the challenge of skepticism threatens human knowledge. The stakes seem even higher with the problem of mental causation, for this problem threatens to take away both agency and cognition.

Three Problems of Mental Causation

What then are the assumptions and presumptions that make trouble for mental causation, prompting us to attempt its 'vindication'? I believe there are three doctrines currently on the scene each of which poses prima facie difficulties for mental causation. The first two have been with us for some time; the third, though not new, has begun to receive serious new considerations. One is 'mental anomalism,' the claim that there are no causal laws about psychological phenomena. The second source of the problem is computationalism and content externalism. The third I call 'causal exclusion.' Each of these generates a distinct problem of mental causation, though the problems are to some extent interconnected. A truly comprehensive theory of mental causation must provide a solution to each problem, a solution that simultaneously satisfies the demands of all three problems.

The Problem of Anomalous Mental Properties

Let us begin with mental anomalism. Davidson's version of this doctrine holds that there are no causal laws (or, in Davidson's terms, 'strict' laws) about psychological phenomena—no such laws connecting mental events with physical events and no such laws connecting mental events with other mental events.[3] But why

does mental anomalism pose a difficulty for mental causation? The initial difficulty arises when anomalism is combined with the widely accepted nomological requirement on causal relations,[4] the condition that events standing in a causal relation must instantiate a causal law. But this seems to make mental causation impossible: mental causation requires mental events to instantiate laws, but mental anomalism says there are no laws about mental events.

Davidson's own proposal is well-known; he calls it 'anomalous monism.' We have already considered it as a mind–body theory and found it wanting; but here our interest lies in Davidson's ingenious argument leading to his physical monism. True, says Davidson, mental events in causal relations must instantiate laws but since there aren't any psychological laws, that can only mean that they instantiate physical laws. This shows that mental events fall under physical kinds (or have true physical descriptions), from which it further follows, argues Davidson, that they are physical events. This is the monism in his anomalous monism. The general upshot of the argument is that for any event to enter into a causal relation, it must be covered by a physical law and hence be part of the physical domain. Causal relations can obtain only between physical events covered by physical laws, although of course some of these events are also mental events. The causal structure of this world—the total set of causal relations that hold in this world—is entirely due to the prevailing physical laws. Mental events are causally efficacious therefore only because they are identical with causally efficacious physical events.

But this ingenious solution has failed to satisfy very many philosophers. On the contrary, there has been an impressive unanimity among Davidson's commentators on just why anomalous monism falls short as an account of mental causation.[5] Take any mental event *m* that stands in a causal relation, say as a cause of event *e*. According to Davidson, this causal relation obtains just in case *m* and *e* instantiate a physical law. Thus *m* falls under a certain physical (perhaps, neural) kind *N*, *e* falls under a physical kind *P*, and an appropriate causal law connects events of kind *N* with events of kind *P*. But this apparently threatens the causal relevance of mentality: the fact that *m* is a mental event—that it is the kind of mental event it is—appears to have no role in determining what causal relations it enters into. Event *m*'s causal relations are fixed, wholly and exclusively, by the totality of its physical properties, and there is in this picture no causal work that *m*'s mental properties can, or need to, contribute.[6] If mental properties were arbitrarily redistributed over the events of this world, or even if mentality were wholly removed from this world—possibilities apparently left open by Davidson's mental anomalism—that would not affect a single causal relation between events of this world, leaving the causal structure of the world entirely untouched. This seems to consign mental properties to the status of epiphenomena.[7] Thus the problem of mental causation arising out of mental anomalism is to answer this question: *How can anomalous properties be causal properties?* A solution to this problem would have to show either that contrary to Davidson, mental properties are not in reality anomalous, or that being anomalous in Davidson's sense is no barrier to their having causal relevance or being causally efficacious.

There have been several attempts to rehabilitate the causal status of mental properties within the constraint of mental anomalism. Most of these attempts have taken the tack of relaxing, or somehow circumventing, the nomological requirement on causal relations. This is usually done in one of three ways. First, you may want to allow laws that are less than 'strict,' perhaps laws tacitly qualified by 'ceteris paribus' clauses, to subsume individual events in causal relations, and argue that there are nonstrict laws of this kind involving mental properties. Second, you look to some form of counterfactual dependency, rather than subsumptive causal laws, to generate causal relations. Fodor's approach[8] is an example of the first strategy; those of LePore and Loewer's[9] and of Horgan's[10] are examples of the second. A third approach (which is consistent with the second) is to define a notion of causal relevance or efficacy weaker than causation regulated by strict laws. A version of this approach, recently embraced by Davidson,[11] attempts to invoke supervenience of the mental on the physical to explain the causal relevance of the mental. But, as we will see, mind–body supervenience itself can be seen to lead to difficulties for mental causation.

The Problem of Extrinsic Mental Properties

Let us begin with syntacticalism, the view that only 'syntactic' properties of mental states, not their 'semantic' (or 'content' or

'representational') properties, can be causally relevant—in particular, to behavior causation.[12] Given the further assumption that the mentality of an important class of mental states, like beliefs and desires, consists in their semantic or representational character, syntacticalism appears to force upon us the conclusion that the intentional properties of mental states, the properties that are constitutive of their mentality, are causally irrelevant. But what persuades us to take syntacticalism seriously?

Syntacticalism most naturally arises in the context of computationalism, an approach that urges us to view mental processes as computational processes on internal representations, on the model of information processing in digital computers. It is apparent that computational processes—that is, causal processes that constitute computation—are sensitive to the syntax, not semantics, of the representations or data structures that are being manipulated; it is the shapes, not meanings, of symbols that determine the course of computation. It matters none to computation whether a given string of 1s and 0s means the inventory count of toothpaste at the local supermarket, the atmospheric pressure in Providence at noon today, the altitude of an airplane on a landing approach, or nothing at all. Similarly, if mental activities are computational processes on beliefs, desires, and such, it would seem that it is the syntactic shapes of these states, not their representational contents, that are causally relevant.[13]

The essential problem here is easily divorced from computationalism and talk of an inner mental language with a syntax and semantics. The internal cause of physical behavior must be supervenient on the total internal state of the agent or organism at the time.[14] For it seems highly plausible to assume that if two organisms are in an identical total internal state at a given time, they will emit identical motor output. However, semantic properties of internal states are not in general supervenient on their *synchronous internal* properties, for as a rule they involve facts about the organism's history and ecological conditions.[15] Thus two organisms whose total states at a given time have identical intrinsic properties can differ in respect of the semantical properties they instantiate; they can differ in the contents of their beliefs and desires, the extensions of their homophonic predicates, and the truth conditions of their homophonic sentences. But prima facie these semantical differences should make no difference to behavior output. The realization that ordinary content ascriptions have this extrinsic/relational dimension is one of the more notable developments in the philosophy of mind and language during the past two decades.[16] You on this earth have the belief that water is wet; yet, as the story goes, your exact physical duplicate on Twin Earth believes that XYZ is wet, not that water is wet. Frogs on the earth, when appropriately stimulated optically, have the 'belief' that a fly is flitting across its visual field (or, at any rate, 'sees' a fly); frogs on another planet without flies, when identically stimulated, don't have a belief about flies, or at any rate are not in a state that represent flies—they 'believe' that a 'schmy' is flitting across its visual field (schmies are tiny black bats which the frogs of this other planet feed on). Thus, that a given intentional state of an organism instantiates a certain semantic property is a *relational* fact, a fact that essentially involves the organism's relationship to various external environmental and historical factors. This makes semantic properties relational, or extrinsic, whereas we expect causative properties involved in behavior production to be nonrelational, or intrinsic, properties of the organism. If inner states are implicated in behavior causation, it seems that all the causal work is done by their 'syntactic,' or at any rate internal/intrinsic, properties, leaving their semantic properties causally otiose. The problem of mental causation generated by syntacticalism therefore is to answer the following question: *How can extrinsic, relational properties be causally efficacious in behavior production?*

So the crux of the problem lies in the supposed fact that mental properties, in particular, content properties (e.g., being a belief that *P*), are relational properties, extrinsic to the organisms instantiating them, whereas we expect the causative properties of behavior to be intrinsic and internal.[17]

The Problem of Causal Exclusion

The third, and final, problem about mental causation I have in mind arises as follows: suppose that we have somehow put together an account of how mental events can be causes of physical events, an account that meets the requirements of the problems of anomalous mental properties and of syntacticalism. Suppose then that mental event *m*, occurring at time *t*, causes physical event *p*, and let us suppose that this causal

relation holds in virtue of the fact that m is an event of mental kind M and p an event of physical kind P. Does p also have a physical cause at t, an event of some physical kind N?

To acknowledge mental event m (occurring at t) as a cause of physical event p but deny that p has a physical cause at t would be a clear violation of the causal closure of the physical domain, a relapse into Cartesian interactionist dualism which mixes physical and nonphysical events in a single causal chain. But to acknowledge that p has also a physical cause, p^*, at t is to invite the question: Given that p has a physical cause p^*, what causal work is left for m to contribute? The physical cause therefore threatens to exclude, and preempt, the mental cause. This is the problem of causal exclusion. The antireductive physicalist who wants to remain a mental realist, therefore, must give an account of how the mental cause and the physical cause of one and the same event are related to each other. Token physicalism, like Davidson's anomalous monism, is not enough, since the question ultimately involves the causal efficacy of mental *properties*, and antireductionism precludes their reductive identification with physical properties. Thus the problem of causal exclusion is to answer this question: *Given that every physical event that has a cause has a physical cause, how is a mental cause also possible?*

These then are the three principal ways in which I believe the problem of mental causation arises in current debates in philosophy of mind. This means that there really are three separable problems, although of course this does not preclude their resolution by a single unified approach. Here I will not deal directly with the first two problems; as I said at the outset of this talk, what I want to do is to develop the third problem—the exclusion problem—in a more concrete and detailed way by focusing on the two theses we discussed in my first lecture, namely the claim that the mental supervenes on the physical and the claim that the mental is realized in the physical. I hope to show how both mind-body supervenience and physical realizationism can be seen to lead to prima facie difficulties for mental causation. In a later lecture I will discuss how physical realizationism, via a functional reduction of mental properties, presents an opening for a possible accommodation of mentality within the causal structure of the physical world, although the opening may well turn out to be not wide enough to let in all mental properties.

The Supervenience Argument, or Descartes' Revenge

In my first lecture I argued that mind–body supervenience could usefully be thought of as defining minimal physicalism—that it is the minimal commitment that anyone who calls herself a physicalist should be willing to accept. We saw also that mind–body supervenience is entailed by physical realizationism, the thesis that mental properties are instantiated in virtue of being realized by physical properties in physical systems. Moreover emergentism, too, is arguably committed to mind–body supervenience: if two systems are wholly alike physically, we should expect the same mental properties to emerge, or fail to emerge, in each.

Let us now turn to an argument designed to show that mind–body supervenience itself leads to apparent difficulties with mental causation. If we take the supervenience thesis to define minimal physicalism, as I earlier suggested, the argument will show that these difficulties will beset physicalism in general—that is, even the weakest form of physicalism must come to terms with this argument one way or another. If this is right, abandoning the substantival dualism of Descartes doesn't get us out of the woods as far as mental causation is concerned. Indeed one notable development in the recent philosophy of mind is the return of the problem of mental causation as a serious challenge to mainstream physicalism, a phenomenon that would have amused Descartes.

I will now proceed to construct a dilemma-style argument that apparently leads to the conclusion that mental causation is unintelligible. In essence the argument to be presented is the result of superimposing mind-body supervenience on the causal exclusion problem. We begin by setting forth the two horns of the dilemma:

(i) Either mind-body supervenience holds or it fails.

But what does mind-body supervenience assert? Let me restate the mind–body supervenience thesis:

> *Mind-body supervenience* Mental properties supervene on physical properties in the sense that if something instantiates any mental property M at t, there is a physical base property P such that the thing has P at t, and necessarily anything with P at a time has M at that time.

Note that a base property is *necessarily* sufficient for the supervenient property; the

necessity involved here is standardly taken to be at least *nomological necessity*—so that if mind–body supervenience holds, it holds in all worlds that share with our world the same fundamental laws of nature.

Returning to (i), we briefly pursue the second horn first:

(ii) If mind-body supervenience fails, there is no visible way of understanding the possibility of mental causation.

According to Jerry Fodor, 'If mind/body supervenience goes, the intelligibility of mental causation goes with it.'[18] To my knowledge he has never explained why he has said this (and not just once!). Fodor is not alone in tying the fate of mental causation to supervenience: Horgan, for example, has argued for the physical supervenience of qualia on the ground that it is needed to make qualia causally efficacious.[19] But what exactly is the connection between supervenience and mental causation? The simplest and most obvious reason for the physicalist to accept (ii) lies, I think, in her commitment to *the causal closure of the physical domain*, an idea that has already made a brief appearance above. One way of stating the principle of physical causal closure is this: If you pick any physical event and trace out its causal ancestry or posterity, that will never take you outside the physical domain. That is, no causal chain will ever cross the boundary between the physical and the nonphysical. The interactionist dualism of Descartes is in clear contravention of this principle. If you reject this principle, you are ipso facto rejecting the in-principle completability of physics—that is, the possibility of a complete and comprehensive physical theory of all physical phenomena. For you would be saying that any complete explanatory theory of the physical domain must invoke nonphysical causal agents. Never mind a complete physical explanation of everything there is; there couldn't even be a complete physical explanation of everything physical. It is safe to assume that no serious physicalist could accept such a prospect.

Now if mind–body supervenience fails—that is, if the mental domain floats freely, unanchored in the physical domain, causation from the mental to the physical would obviously breach the physical causal closure. Mind–body supervenience grounds each mental phenomenon in the physical domain by providing for it a set of physical conditions that are (at least) nomologically sufficient for it and on which its occurrence depends. A corollary is the thesis that no mental phenomenon can occur, and no mental property can be instantiated, unless an appropriate physical base condition is present. Every mental event, be it a sensation like pain or itch, or an intentional state like belief and desire, must have a physical basis: it occurs because an appropriate physical basis is present, and it would not occur if such a basis was absent.[20] These comments hold true if you wish to speak in terms of realization. If any mental property gets instantiated because, and only because, one of its physical realizers is instantiated, there is a similar dependence of mental occurrences on physical occurrences.

In any case mind–body supervenience brings mental phenomena within the ambit of the physical: the physical determines the mental, and in that sense the mental does not constitute an ontologically independent domain that injects causal influences into the physical domain from the outside. Now it is another question whether or not mind-body supervenience brings the mental *close enough* to the physical to allow mental causation to circumvent the constraint of the physical causal closure.[21] But we can skirt this question here, for if the answer is no, that would only show that mind–body supervenience isn't enough to give us a solution to the problem of causal exclusion of the mental by the physical. But there is a potentially more serious problem with supervenience: mind-body supervenience may itself be a source of the problem. That is, mind-body supervenience, far from being part of the solution, as hoped for by Fodor, Horgan, and others, may turn out to be part of the problem. Let us now look into this possibility.

(iii) Suppose that an instance of mental property M causes another mental property M^* to be instantiated.

So this is a case of mental-to-mental causation, one in which an instance of a mental property causes an instance of another mental property. We may take 'instances' or 'instantiations' of properties as events, states, or phenomena. For brevity, I will often speak of one property causing another property; this is to be understood to mean that an *instance* of the first causes an *instance* of the second.[22] Returning to our argument, we see that (ii), the supervenience premise, yields:

(iv) M^* has a physical supervenience base P^*.

We now ask the following critical question: *Where does this instance of M* come from? How does M* get instantiated on this occasion?* There apparently are two possible answers to consider:

(v) *M** is instantiated on this occasion: (a) because, ex hypothesi, *M* caused *M** to be instantiated; (b) because *P**, the physical supervenience base of *M**, is instantiated on this occasion.

I hope that you are like me in seeing a real tension between these two answers: Under the assumption of mind–body supervenience, *M** occurs because its supervenience base *P** occurs, and as long as *P** occurs, *M** must occur no matter what other events preceded this instance of *M**—in particular, regardless of whether or not an instance of *M* preceded it. This puts the claim of *M* to be a cause of *M** in jeopardy: *P** alone seems fully responsible for, and capable of accounting for, the occurrence of *M**.[23] As long as *P**, or another base property of *M**, is present, that absolutely guarantees the presence of *M**, and unless such a base is there on this occasion, *M** can't be there either. Given this, the only way anything can have a role in the causation of *M** would have to be via its relationship to *M**'s supervenience base *P**, and as far as I can see, the only way of reconciling the claim of *M* to be a cause of *M** with the fact that *M** has *P** as its supervenience base is to accept this:

(vi) *M* caused *M** *by causing P**. That is how this instance of *M* caused *M** to be instantiated on this occasion.

There may be a plausible general principle involved here, which is by itself sufficient to justify (vi) even if you do not see the tension in (v), and it is this: *To cause a supervenient property to be instantiated, you must cause its base property (or one of its base properties) to be instantiated.* To relieve a headache, you take aspirin: that is, you causally intervene in the brain process on which the headache supervenes. That's the only way we can do anything about our headaches. To make your painting more beautiful, more expressive, or more dramatic, you must do physical work on the painting and thereby alter the physical supervenience base of the aesthetic properties you want to improve. There is no direct way of making your painting more beautiful or less beautiful; you must change it physically if you want to change it aesthetically—there is no other way.

But note what (vi) asserts: it says that a mental property *M* causes a physical property *P** to be instantiated. This of course is a case of mental-to-physical causation. So what our argument has shown so far is this: *Under the mind-body supervenience assumption, mental-to-mental causation implies, or presupposes, mental-to-physical causation.* So the question that we now face is whether we can make sense of mental-to-physical causation—that is, under the premise of mind–body supervenience.[24]

Going back to (vi): we see that on the assumption of mind–body supervenience, it follows:

(vii) *M* itself has a physical supervenience base *P*.

We must now compare *M* and *P* in regard to their causal status with respect to *P**. When we reflect on this point, I believe, we begin to see reasons for taking *P* as preempting the claim of *M* as a cause of *P**. If you take causation as grounded in nomological sufficiency, *P* qualifies as a cause of P*, for, since *P* is sufficient for *M* and *M* is sufficient for *P**, *P* is sufficient for *P**. If you choose to understand causation in terms of counterfactuals, again there is good reason to think that *P* qualifies: if *P* hadn't occurred *M* would not have occurred (we may assume, without prejudice, that no alternative physical base of *M* would have been available on this occasion), and given that if *M* had not occurred *P** would not have occurred, we may reasonably conclude that if *P* had not occurred, *P** would not have either.[25]

It seems then that we are now blessed with an overabundance of causes: both *M* and *P* seem severally eligible as a sufficient cause of *P**. And it is not possible to escape the threat of causal overdetermination by thinking of the situation as involving a causal chain from *P* to *M* and then to *P**, with *M* as an intermediate causal link. For the relation from *P* to *M* is not happily thought of as a causal relation; in general, the relation between base properties and supervenient properties is not happily construed as causal.[26] For one thing, the instantiations of the related properties are wholly simultaneous, whereas causes are standardly thought to precede their effects; second, it is difficult, perhaps incoherent, to imagine a causal chain, with intermediate links, between the subvenient and the supervenient properties. What intermediate stages could link the beauty of a painting

to its physical properties? What intermediary events could causally connect a mental event with its subvenient physical base? Would such intermediaries themselves be mental or physical? Moreover, for the present case, the causal chain approach, in taking *M* to be a nonphysical cause of *P**, would violate the causal closure of the physical domain, an option foreclosed to the physicalist.

Nor does it seem plausible to take *M* and *P* together to constitute a single sufficient cause of *P**. There are two reasons for this. First, *P* alone is causally sufficient for *P**, and so is *M*. It is difficult to see how *M* and *P* together can pack any more causal power than *M* alone or *P* alone. Second, this approach is plausible only if it claims *M* to be a necessary component in the causation of *P**, and this means that, as with the causal chain proposal, it involves a violation of the physical causal closure. For a complete causal explanation of why *P** was instantiated on this occasion would have to advert to the presence of a nonphysical causal agent, *M*, in addition to *P*.

And, finally, it is not possible to take this simply as a case of causal overdetermination—that the instance of *P** is causally overdetermined by two sufficient causes, *P* and *M*. Apart from the implausible consequence that it makes every case of mental causation a case of overdetermination, this approach encounters two difficulties: first, in making a physical cause available to substitute for every mental cause, it appears to make mental causes dispensable in any case; second, the approach may come into conflict with the physical causal closure. For consider a world in which the physical cause does not occur and which in other respects is as much like our world as possible. The overdetermination approach says that in such a world, the mental cause causes a physical event—namely that the principle of causal closure of the physical domain no longer holds. I do not think we can accept this consequence: that a minimal counterfactual supposition like that can lead to a major change in the world.

It seems to me that the most natural way of viewing the situation is this:

(viii) *P* caused *P**, and *M* supervenes on *P* and *M** supervenes on *P**.

This explains the observed regularities between *M*-instances and *M**-instances, and those between *M*-instances and *P**-instances.[27] These regularities are by no means accidental; in a clear sense they are law-based, and may even be able to support appropriate counterfactuals. However, if we understand the difference between genuine, productive and generative causal processes, on the one hand, and the noncausal regularities that are observed because they are parasitic on real causal processes, we are in a position to understand the picture recommended by (viii). In the case of supposed *M*-*M** causation, the situation is rather like a series of shadows cast by a moving car: there is no causal connection between the shadow of the car at one instant and its shadow an instant later, each being an effect of the moving car. The moving car represents a genuine causal process, but the series of shadows it casts, however regular and lawlike it may be, does not constitute a causal process.[28] Hence we have:

(ix) The *M*-to-*M** and *M*-to-*P** causal relations are only apparent, arising out of a genuine causal process from *P* to *P**.

Whence a dilemma:

(x) If mind–body supervenience fails, mental causation is unintelligible; if it holds, mental causation is again unintelligible. Hence mental causation is unintelligible.

That then is the supervenience argument against mental causation, or Descartes' revenge against the physicalists. I believe it poses a serious challenge to physicalism by casting doubts on the possibility of mental causation within the parameters it sets for itself. Descartes' difficulties arose from the duality of mental and material substances. Current mainstream physicalism, which calls itself 'nonreductive physicalism,' runs into parallel difficulties on account of its commitment to the duality of psychological and physical properties—or its failure to make a reductionist commitment for psychological properties. For it is clear that the tacit assumption that gets the supervenience argument going is mind–body antireductionism; if the mental properties are viewed as reducible to physical properties in an appropriate way, we should expect to be able to disarm the argument (although of course the details will need to be worked out).

One good question to raise about the foregoing argument is this: Wouldn't the same argument show that all properties that supervene on basic physical properties are epiphenomenal, and that their causal efficacy is unintelligible? However, there seems to be more than ample reason to think that geological properties, say,

are supervenient on fundamental physical properties, and if mind–body supervenience could be shown to put mental causation in jeopardy, wouldn't the very same considerations do the same for geological properties? But no one seems to worry about geological causation, and there evidently seems no reason to start worrying. If so, shouldn't we conclude that there must be something wrong with the argument of the preceding section?[29]

I will deal with this question in detail in my two remaining lectures. As I see it, however, the heart of the issue here is this: with properties like geological and biological properties, we are much more willing, intuitively, to accept a reductionist picture in relation to basic physical properties. I believe that this is true even for philosophers who are vocal in their claim that antireductionism holds across the board, at all levels in relation to their lower levels, and that geological and biological properties are no more reducible to basic physical properties than mental properties. Clearly it is possible that their antireductionism is more correct about mental properties than about these other 'higher-level' physical properties. . . .

NOTES

1. For Pierre Gassendi's vigorous challenge to Descartes, see René Descartes, *The Philosophical Writings of Descartes*, vol. 2, John Cottingham, Robert Stoothoff, and Dugald Murdoch, eds., (Cambridge: Cambridge University Press, 1985), p. 238.
2. "Making Mind Matter More," reprinted in *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990), p. 156.
3. Donald Davidson, "Mental Events," reprinted in *Essays on Actions and Events* (Oxford: Oxford University Press, 1980). For wholly different considerations in favor of mental anomalism, see Norman Malcolm, *Memory and Mind* (Ithaca, NY: Cornell University Press, 1977), and Bruce Goldberg, "The Correspondence Hypothesis," *Philosophical Review 77* (1968): pp. 439–54.
4. This condition is not as widely accepted as it used to be. All known alternatives have their own difficulties, however, and it is fair to say that the nomological conception of causation, in its many variants, is still 'the received view.'
5. To cite a few of the papers in which this issue has been raised, Frederick Stoutland, "Oblique Causation and Reasons for Action," *Synthese* 43 (1980): pp. 351–67; Ted Honderich, "The Argument for Anomalous Monism," *Analysis* 42 (1982): pp. 59–64; Ernest Sosa, "Mind-Body Interaction and Supervenient Causation," *Midwest Studies in Philosophy* 9 (1984): pp. 271–81; Jaegwon Kim, "Self-understanding and Rationalizing Explanations," *Philosophia Naturalis* 82 (1984): pp. 309–20; Louise Antony, "Anomalous Monism and the Problem of Explanatory Force," *Philosophical Review* 98 (1989): pp. 153–87. Davidson defends his position in "Thinking Causes," in *Mental Causation*, John Heil and Alfred Mele, eds., (Oxford: Clarendon, 1993). This volume also includes rejoinders to Davidson by Kim, Sosa, and Brian McLaughlin.
6. This remains true even if Davidson's 'strict law' requirement on causation is weakened so that nonstrict laws (or *ceteris paribus* laws)—including nonstrict psychophysical laws—are allowed to support causal relations. For suppose that *m* falls under mental kind *M* and that there is a nonstrict law connecting *M* with *P* (or another physical kind *P** under which *p* falls). Might this show *M* to be efficacious in *m*'s causation of *p?* Hardly, for given that *m*'s causation of *p* is covered by the strict law connecting *N* and *P*, what *further* causal work is left for *M*, or the law connecting *M* and *P?* This is a form of 'the exclusion problem'; see below for further discussion.
7. Brian McLaughlin calls this 'type epiphenomenalism' in "Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical," *Philosophical Perspectives* 3 (1989): pp. 109–35.
8. Jerry A. Fodor, "Making Mind Matter More," *Philosophical Topics* 17 (1989): pp. 59–80. In his "Thinking Causes" (in *Mental Causation*), Davidson seems to buy into Fodor's proposal. To see why this won't work, at least for Davidson, see note 6 above.
9. Ernest LePore and Barry Loewer, "Mind Matters," *Journal of Philosophy* 93 (1987): pp. 630–42.
10. Terence Horgan, "Mental Quausation," *Philosophical Perspectives*, 3 (1989): pp. 47–76.
11. Davidson, "Thinking Causes," in *Mental Causation*. For an earlier attempt to make use of supervenience to explain mental causation, see my "Epiphenomenal and Supervenient Causation," *Midwest Studies in Philosophy* 9 (1984): pp. 257–270, reprinted in *Supervenience and Mind*. I explain why I now think this approach to be inadequate in *Supervenience and Mind*, pp. 358–62.
12. See Stephen P. Stich, *From Folk Psychology to Cognitive Science* (Cambridge: MIT Press, 1983).
13. For a clear development of these issues, see Ned Block, "Can Mind Change the World?" in *Meaning and Method*, George Boolos, ed., (Cambridge: Cambridge University Press, 1990).
14. For a more detailed statement of this argument, see Stephen P. Stich, "Autonomous Psychology and the Belief-Desire Thesis," *The Monist* 61 (1978): pp. 573–91.
15. There are well-known considerations supporting a view of this kind; see, for example, Hilary Putnam, "The Meaning of 'Meaning,'" in *Philosophical Papers*, vol. 2 (Cambridge: Cambridge University Press, 1975); Tyler Burge, "Individualism and the Mental," *Midwest Studies in Philosophy* 4 (1979): pp. 73–121; Stich, "Autonomous Psychology and the Belief-Desire Thesis"; Kim, "Psychophysical Supervenience," *Philosophical Studies* 41 (1982): pp. 51–70.
16. Due to the works by Hilary Putnam, Saul Kripke, Tyler Burge, and others.

17. For instructive and helpful discussion of issues concerning the causal/explanatory efficacy of contentful mental states, see Lynne Rudder Baker, *Explaining Attitudes* (Cambridge, UK: Cambridge University Press, 1995), and Pierre Jacob, *What Minds Can Do* (Cambridge, UK: Cambridge University Press, 1997).
18. Jerry Fodor, *Psychosemantics* (Cambridge, MA: MIT Press, 1987), p. 42.
19. Terence Horgan, "Supervenient Qualia," *Philosophical Review* 96 (1987): pp. 491–520.
20. On content externalism, wide-content states will not supervene on internal physical properties of the subject, but physicalists will not deny that they supervene on the subject's extrinsic/relational physical properties. For the present paper, we will ignore the issues that arise from content externalism. But see the works by Baker and Jacob cited in note 17.
21. On this issue see my "Postscripts on Mental Causation," in *Supervenience and Mind* (Cambridge, UK: Cambridge University Press, 1993).
22. Strictly speaking, this doesn't go far enough: it must further be the case that one instance causes another instance *in virtue of the fact that the first is an F-instance and the second is a G-instance.*
23. This argument is based on what I have called 'the principle of causal/explanatory exclusion'; see, for example, my "Mechanism, Purpose, and Explanatory Exclusion," reprinted in *Supervenience and Mind.*
24. We could have begun with (vi) as our initial premise of mental causation. The point of starting with (iii) is to show that the argument applies to mental-mental causation as well as to mental-physical causation. On the assumption of mind-body supervenience, the former is as problematic, in my view, as the latter.
25. This of course is not to assume transitivity for counterfactuals in general.
26. One philosopher who holds the unorthodox view that the base properties 'cause' the supervenient properties is John Searle, in his *The Rediscovery of the Mind* (Cambridge: MIT Press, 1992).
27. Note, however, that these regularities are likely to be restricted in generality. The reason is that M's alternative supervenience bases cannot be counted on to cause P^* and hence M^*.
28. On the distinction between 'causal process' and 'pseudo-process,' see Wesley Salmon, *Scientific Explanation and the Causal Structure of the World* (Princeton, NJ: Princeton University Press, 1984).
29. Several philosophers have raised exactly these questions (though not necessarily directed against our first argument); for example, Lynne Rudder Baker, "Metaphysics and Mental Causation," in *Mental Causation*; Robert Van Gulick, "Three Bad Arguments for Intentional Property Epiphenomenalism," *Erkenntnis* 36 (1992); Louise M. Antony, "The Inadequacy of Anomalous Monism as a Realist Theory of Mind," in *Language, Mind, and Epistemology*, G. Preyer, F. Siebelt, and A. Ulfig, eds., (Dordrecht: Kluwer, 1994).

Emergentisms, Ancient and Modern

Jonardon Ganeri

1. Emergence: the core issues

As a position in contemporary philosophy of mind, emergentism developed out of the early work of a number of British philosophers.[1] According to British emergentism, each special science (chemistry, biology, psychology, and so on) describes a range of causal powers that emerge from but are irreducible to the causal powers of physical particles:

> British emergentism maintains that some special science kinds from each special science can be wholly composed of types of structures of material particles that endow the kinds in question with fundamental causal powers. Chemical elements, in virtue of their minute internal structures, have the power to bond with certain others. Certain biological organisms, in virtue of their minute internal structures, have the powers to breathe, to digest food, and to reproduce. *And certain kinds of organisms, in virtue of the minute internal structures of their nervous systems, have 'the power of cognizing, the power of being affected by past experiences, the power of association, and so on'* (Broad 1925, p. 436). The property of having a certain type of structure will thus endow a special science kind with emergent causal powers. Such a structure will have an emergent causal power as a matter of law, but the law will not be 'reducible to' or 'derivative from' laws governing lower levels of complexity and any boundary conditions involving the arrangements of particles. (McLaughlin 1992, pp. 50–1; my italics)

Originally published in *Mind*, 120 (479), July 2011.

Generalising, the satisfaction of two conditions is typically regarded as necessary for any theory to be emergentist:

(1) Mental properties *supervene* on physical properties.[2]

(2) Mental properties confer on their instances causal powers *irreducible* to the causal powers conferred by the properties supervened on.[3]

One idea, indeed the leading one in discussions about emergence, is that systems of appropriate organizational complexity have causal powers which the components in the system, whether individually or together, do not. Emergentism hopes to give sense to the idea that mental properties are metaphysically *dependent* on physical properties but yet possess causal *autonomy* with respect to them.[4]

Jaegwon Kim (Kim 2006a; cf. also Kim 2006b) agrees that any version of emergentism is committed to a supervenience thesis and an irreducibility thesis, and specifically that the core emergentist idea that emergent properties contribute new causal powers neither explicable by nor predictable from the basal properties is a denial of functional reducibility. The two key issues for the development of emergentism as a viable theory, he argues, are (i) to give *a positive characterization* of the relation of emergence, beyond the mere denial of reducibility; and (ii) to solve *the problem of downward causation*, otherwise known as the exclusion problem or the supervenience problem. This is the problem that an instantiation of the supervenience base is apparently a sufficient cause for any effect attributed to an instantiation of the supervening properties. One seems forced to choose between reductionism (mental properties are 'nothing but' physical properties) and epiphenomenalism (mental properties are distinct from physical properties, but the residue is inefficacious): genuinely novel emergent causal power is excluded.

As I will demonstrate here, the ancient philosopher of mind Bṛhaspati[5] and at least some of his successors are emergentists. Responsive to the key issues Kim has identified, their work contains materials for the articulation of a conception of the mind's dependence on, and autonomy from, the physical, one that will be of considerable interest to contemporary philosophers of mind.

2. Indian physicalism

There are references in the Indian texts to a group of renegade free-thinkers whose views about human life are radically at odds with then-prevailing belief. These worldly intellectuals deny the existence of anything that smacks of the supernatural—such as transcendental beings, immaterial souls, or heavenly other-worlds. Life, they say, is for living here on earth. And they have a most interesting account of what human life itself consists in. A human person is a material body, made, like everything else, out of the four elements—but one in which thought, reason, intelligence, and consciousness arise as the physical elements are transformed, in a way similar to the way that the process of fermentation leads to the emergence of the power to intoxicate in a mixture of appropriate ingredients. The views of these philosophers, who were known as Lokāyata ('worldly'), or more commonly Cārvāka, and whose central figure is the enigmatic Bṛhaspati, have been deeply unfashionable, their treatises destroyed or left to rot, their ideas subject to fierce and hostile criticism. That they were nevertheless still known in the sixteenth century is evinced by the report of Abu al-Faḍl, who describes their theory for the benefit of the Mughal Emperor Akbar, saying that 'They regard paradise as a state in which man lives as he chooses, free from the control of another, and hell the state in which he lives subject to another's rule . . . They admit only of such sciences as tend to the promotion of what is external, that is, a knowledge of just administration and benevolent government' (al-Faḍl 1873–1907: vol. 3, p. 217). No truck is given here to religion and other 'inner' spiritual disciplines.

The contemporary inquiry into the foundations of naturalism gives us new reasons to examine the views of these thinkers. Their most important contribution, I will claim, is a distinctive interpretation of the doctrine that psychological states are emergent physical states. A separable claim is that the self is identical to the physical body. This second claim, which I am not going to consider here, has to do with the material constitution of the self and its identity over time, and the view is a version of Animalism, the claim being that a person is identical with the human animal and not with either an immaterial soul or a psychological continuum.

According to Bṛhaspati, thinking is due to the four constitutive principles of matter, just as the power to intoxicate is due to the ingredients in the wine. What we call a human body, or a sense organ, or a physical object, is just a combination (*samudaya*; an assemblage) of earth, fire, air, and water; indeed, these four kinds of matter are all there is. A person is a human body endued with thinking, and individual lives differ one from another as bubbles differ in water. Recent work on the reconstruction of Bṛhaspati's text allows us to conjecture that it begins as follows:

1.1 Next then we will examine the nature of the reals.

1.2 Earth, fire, air, and water are the reals.

1.3 Their combination is called the 'body,' 'senses,' and 'objects.'

1.4 Consciousness (*caitanya*) [is formed] out of these [elementary reals].

1.5 As the power to intoxicate [is formed] out of fermenting ingredients.

1.6 A human being (*puruṣa*) is a body qualified by consciousness.

1.7 [Thinking is] from the body alone.

1.8 Because of its presence when there is a body.[6]

Two initial observations: First, Bṛhaspati's commitment to physicalism seems to be unambiguous. He says that earth, water, fire, and air are real, nothing else, and that what we call an object, a body, or a sense-organ is just an aggregation. The science of the four elemental reals just is (or rather, was)[7] the science of physics, and everything which exists, it is asserted, is identical to the elements or to some combination of them. Second, Bṛhaspati's commitment to the completeness of physics is evident in his further claim that all variation in the world is due to variation in 'origin' (*janma*). The varied patterns which are seen in the eye of a peacock's tail feathers come about as a result of details in their provenance, and the same explanation works for all other worldly variation:

2.1 The world is varied due to variations in origin.

2.2 As the eye in the peacock's tail.[8]

I take this to mean that there is a complete physical causal history for every change or difference, that is, as a version of the thesis that every physical effect (every 'variation') is determined by antecedent physical causes (its 'origin').

Bṛhaspati's philosophy of mind can be resolved into a pair of theses:

(Thesis 1) A human being consists in a living body made out of the four elements which, in that combination, instantiates mental properties.

It is striking that the term 'self' is not used here at at all, but only the term 'human being' (*puruṣa*).[9] The difficulty is to extract further resolution from a second thesis:

(Thesis 2) It is *due to* the combination of the elements in the body that mental properties are instantiated.

The trouble is with the ablative, which I have translated, as neutrally as possible, as 'due to.' Is the claim that thinking consists in the elements combined in a certain way, in other words, that it is *made from* them (an ablative of composition); or that it is the claim that *because of* the elements there is thinking (an ablative of explanation); or is it the claim that thinking is *produced out of* them (an ablative of causation)? Later sources will disambiguate this ablative in two different ways, as well as offering a distinct three-way disambiguation of the statements taken as a group. These disambiguations generate a range of philosophical positions about the mind-body problem. I will argue that from among the ensuing positions one can retrieve materials for a distinctive variety of emergentism.

3. Epiphenomenalism: Bṛhaspati and Dichaearchus

Among these various possibilities one suggestion is that Cārvāka philosophy of mind is a form of epiphenomenalism. According to this possibility the mind is a by-product of the material body, lacking in causal powers of its own. The question has a long history, and indeed goes back to the first presentation of Bṛhaspati's thought to a European audience. In a famous and widely circulated lecture given to a public meeting of the Royal Asiatic Society in London, February 3, 1827, Henry T. Colebrooke conjectured of Bṛhaspati that '[a]mong the Greeks, Dicaearchus of Messene held the same tenet' (Colebrooke 1837, p. 429). Dicaearchus (c. 350–285 BCE) was a member

of the Lyceum and a defender of the *harmonia* theory, put forward by Simmias in the *Phaedo*, that the soul is a 'tuning' (*harmonia*) or 'tempering/blend' (*krasis*) of the body, that a blend of hot, cold, fluid, and dry material is to the soul what the tuning is to the lyre (*Phaedo* 86b7–c2). Dicaearchus wrote a dialogue, now lost, about the soul, which is mentioned by several later authors. One important source is Cicero, who reports:

> In the remaining two books, [Dicaearchus] introduces a certain Pherecrates, an old man from Phthia, said to be a descendant of Deucalion, who maintains the following. The soul is nothing at all and this name completely vacuous—animals and animate things are so-called in vain [*anima* meaning 'soul'], for there is neither soul nor spirit in either man or beast. That whole power by which we act or are aware extends evenly through all living bodies and is not separable from the body. In fact, [that power] is nothing, nor is there anything else, apart from the body just alone by itself, so configured that it lives and is aware by the tempering of its nature. (*Tusc. disp.* 1.10.21 [Cicero 1927])

This is certainly the source relied on by Colebrooke, for he describes the tenet in question, the one which he finds also in Bṛhaspati, as being 'that there is no such thing as soul in man; that the principle, by which he perceives and acts, is diffused through the body, is inseparable from it, and terminates with it' (Colebrooke 1837, p. 430). The view which Sextus Empiricus attributes to Dicaearchus is that thinking is 'nothing apart from the body disposed in a certain way' (*Adv. math.* 7.349 [Sextus Empiricus 1935]). Plutarch introduces a very similar view, without attributing it: 'Or is this the case? Namely, that the substance of the soul isn't anything at all; rather, it is the tempered body which possesses the power of thinking and living' (*Against Colotes* 1119ab [Plutarch 1967]).

The analogy with the tuning of a musical instrument is helpful because it reminds us that there are three different things we must keep apart: (i) the blend itself; (ii) the dispositional properties and causal powers that the body has, and for which the blend is the categorical base; and (iii) the effects of the blend either on the body or on other things. One might think of a block of ice, with crystalline structure, brittleness, and a capacity to cool other things. Reviewing ideas about the *harmonia* theory, Victor Caston observes that the idea was understood by Dicaearchus as a claim that the soul just *is* the blend (Caston 1997, 2001). Dicaearchus is described as holding that the soul is an attunement of the elements of matter that comprise the body, rather than as a power ascribable to the body in virtue of the attunement.[10] Denying that the attunement has any causal powers of its own, Dicaearchus is an epiphenomenalist.[11]

Certain classical Indian thinkers likewise interpret Bṛhaspati as identifying the mind with a combination of elements in the body (so that it has no causal power over and above that of the body), making him an epiphenomenalist like Dicaearchus, while others claim that his view is that the mind is a distinct power which emerges from the combination but is not identical to it. There is textual evidence of this disagreement among the Indian materialist philosophers of mind. Referring to the basic thesis, that it is due to the physical elements that there is thinking, we are informed that:

> Here, some commentators explain that thinking arises from (*utpadyate*) the elements, while others say that it is made manifest (*abhivyajyate*) [by them]. (Kamalaśīla 1968, pp. 633,15–634,1)

Again:

> Some people restore the connecting verb [in 'due to the elements, thinking'] with 'is manifested' (*abhivyajyate*), but others with 'comes into being' (*prādhurbhavati*). (Prabhācandra 1991, p. 342, 2–3)

This second source goes into much greater detail than any of the others in explaining the concept 'manifestation' in play here. He tells us that a manifestation is something which 'puts together well' or 'refines' and 'perfects' (*saṃskāraka*) what is already there, rather than bringing into being something that was not there before (Prabhācandra 1990, p. 226, 12–13). As such, the manifestation is not a separate thing, over and above the four elements, even though it does have a distinctive characteristic of its own (1990, p. 225, 25); it is not a 'distinct reality' (*tattvāntara*) (1990, p. 115, 13). In this, there is certainly an affinity with the Greek word *harmonia*, which 'derives from a verb for "fitting together", for joining things so as to adapt or accommodate them to each other,' such that tempering 'is the balancing of one against another so as to produce a dynamic whole' (Caston 1997, pp. 321–2). So, the manifestation account, in thinking of states of mind as refinements of the body, is a close cousin of the *harmonia* theory as that theory was understood by Dicaearchus.

Other Aristotelians, notably Galen and Alexander, are drawn to a different reading of the *harmonia* theory, that the *harmonia* gives rise to new causal powers—and our source too distinguishes such a view among the possible interpretations of Bṛhaspati's claim that consciousness arises from the physical elements in the right combination. This same source indeed provides us with a helpful three-fold classification of materialist solutions to the mind-body problem. A materialist must claim that the lack of distinction between mind and body consists in either (i) their necessary identity (*svabhāva*), or (ii) mind being a quality or state (*guṇa*) of body, or (iii) mind being an effect (*kārya*) of the body (Prabhācandra 1990, p. 120, 22-23). This division might be brought into correspondence with the one we have seen in connection with the *harmonia* theory: the mind is either identical to the tempered body, or to the tempering itself, or to a power caused by that tempering. It might also be said to correspond with the modern distinction between reductionism, epiphenomenalism, and emergentism. From here I am interested in the first and third possibilities.

4. From covariance to material causation

The notion of supervenience is explicitly formulated in the Indian discussion of materialism, particularly in critics' descriptions of what materialism is committed to. The supervenience claim is that fixing the body's physical state fixes its mental state: two bodies cannot be distinguishable in terms of their mental properties and yet be indistinguishable in terms of their physical properties. The Latin term *supervenire* is a rendering of the Greek *epi-ginesthai* and *ginesthai epi*, terms which are used in a sense close to the modern one by Alexander and Philoponus. Philoponus in particular uses the notion in contrast with the idea that psychological characteristics simply result from (*apotelesma*) and follow (*hepesthai*) the blend of chemical ingredients, and in such a way as to allow mental states to react back on the body.[12] Donald Davidson was the first contemporary philosopher to promote the use of the notion. He did so as follows:

> Mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events exactly alike in all physical respects but differing in some mental respects, *or that an object cannot alter in some mental respects without altering in some physical respects*. (Davidson 1980, p. 214; my italics)

We can see supervenience as having two components: *dependence* ('nothing can have mental-properties unless it also has physical-properties'), and *determination* ('nothing can be just like a given thing as regards its physical-properties without also being just like it as regards its mental-properties'). In short, 'every mental-property, some physical-property,' and 'same physical-properties, same mental-properties' (Van Cleve 1990, p. 221).

Supervenience, I have claimed, is explicit in the formulations we have of Indian physicalist philosophy of mind. It is not to be found in what one might think of as the obvious place, however. The obvious place is a certain standard argument for physicalism, one at which Bṛhaspati hints in his laconic formulae 1.7–1.8. Other sources present the argument in similar terms:

> Thinking is a quality of the body, because it is present when there is a body and absent when there is none.[13]

I will call this the *covariance argument*. Its premise is that there is a relation of 'presence and absence' between states of the body and states of the mind, and its conclusion is that mental states are states of the body. This relation has two components, 'covariance in presence' (*anvaya*) and 'covariance in absence' (*vyatireka*). The basic pattern, as Cardona (1967–68; 1981) has shown, is:

Covariance in presence:	When B occurs (*tadbhāve*), A occurs (*tadbhāvāt*).
Covariance in absence:	When B is absent (*tadabhāve*), A is absent (*tadabhāvāt*).

Here, A = mind or mental property, and B = body or physical property. On scrutiny, it is clear that we do not yet have a supervenience relation. The two halves of the rule of presence and absence resolve themselves as follows:

[Presence:]	(Necessarily,) anything which has certain physical properties is thinking.
[Absence:]	(Necessarily,) anything which is thinking is a physical thing (i.e. if *x* does not have physical properties, then *x* does not have any mental properties).

It is clear that this does not specify a relation of supervenience. If this is all there is to covariance, then the covariance of the physical with the mental and the mental with the physical lacks the determination component of the supervenience definition. It does not have 'same B-properties, same A-properties' feature, but only the weaker 'every A-property, some B-property.' More particularly, while supervenience is an asymmetric relation between A and B, covariance is entirely symmetric. The covariance argument is at best enthymematic, but it is opaque what suppressed premiss is in the background of appeals to this argument.

Some of our sources, however, introduce the new thought that the body is the 'material cause' of thinking (*upādāna-kāraṇa*), and it turns out not only that the operative notion of 'material cause' does imply supervenience, but that this is made explicit. The idea is that, just as a sculptor could not change the features of a statue without making changes to the material out of which it is made, so too one cannot alter mental states without there being some alteration in their physical basis: we would now call this a relation of constitution. Our sources tell us that that it is part of the notion of material cause that alterations in the material cause are implied by alterations in that which it is the material cause of. In other words, the idea of a material cause carries with it the idea of a supervenience base. Having established this principle, the critics of materialism go on to argue by *reductio* that thinking does not have the physical body as its material cause: if it did, then the mental would supervene on the physical, but it does not. Here is one important text:

> Is it possible that the elements of matter be the 'cause' of thinking, either as the material cause or as a co-operating cause? Certainly not as the material cause, because even when they alter [thinking] does not. If one thing does not alter when another does, that other is not its material cause; the relation between a horse and a cow illustrates this. Thinking is not altered when the material elements that have been transformed into a body alter. This is not [merely] an undemonstrated assertion, for it is well known that thinking which is otherwise engaged is unaltered even by the stab of a knife, which feels [to the preoccupied thinker] no different from a rub of sandal-paste. *In exactly the same way, there can be alterations in thinking without alterations in the [elements comprising the body].* This too is not an unfounded claim, since the joyful emotion one feels when near to a lovely woman alters without one's body changing state. (Prabhācandra 1991, p. 344, 9–15; my italics)

Another source is if anything even clearer:

> Nor is the 'material cause' view correct. For it is well known that the particular cause regarded as the material cause is one such that an alteration in the effect is impossible unless one brings about an alteration in it . . . That is why someone who wants to alter something alters it only by altering its material cause, and in no other way. For when the material cause is present and its power is unimpeded, nobody can prevent the occurrence of its subsequent effect. (Kamalaśīa 1968, pp. 642, 23–43, 5)

What these passages show is that when someone claims that the relation between mind and body is one of 'material causation,' better described as a relation of constitution, that is indeed to make a supervenience claim.

A modal operator is used explicitly in these two formulations, and we are in a position to consider whether the supervenience involved is strong or weak.[14] Inverting the conditional, the claim here is that if x is the material cause of y then it is not possible to bring about an alteration in y without an alteration in x. What has been said is that if x undergoes an alteration, then it is *impossible* to prevent the alteration in y.[15] So the force of the statement is that there are no circumstances in which the intended alteration in y does not occur and yet the alteration in x does; that is, necessarily, if x undergoes alteration G, then y undergoes alteration F. Presuming that the entire claim has the status of a 'rule' or 'law' (*niyama*), and so that there is a second, wide scope, necessity, we can conclude that what is being attributed to the materialist who sees the relation between mind and body as one of 'material causation' is a strong supervenience thesis.[16] Our source, not himself a materialist, points out that to deny strong supervenience of mind on body is not to commit oneself to denying that there are any circumstances in which physical changes necessitate alterations of the mind.

What we have so far established is that the classical theory endorses supervenience, the first requirement for a theory of emergence. With regard to the second requirement, Mark Bedau argues that we should distinguish between *strong* and *weak* conceptions of emergence in the following way.[17] Strong emergence involves 'a requirement that emergent properties are supervenient properties with irreducible causal powers' (Bedau 2003, p. 158; cf.

O'Connor 1994). Weak emergence involves a less demanding requirement, which in Bedau's account is the requirement that 'emergent properties can be derived from micro-level information but only in a certain complex way.' The complexity requirement is what distinguishes weak emergent causal powers from the resultant properties of the system: one cannot *deduce* weakly emergent phenomena from one's knowledge of the basal conditions, but only *simulate* them.[18] Weak emergence uses dynamical systems theory to demonstrate how systems can come to present emergent properties without the strong requirement. The worry is that if mental properties are only weakly emergent, then they will be epiphenomenal. In the next section, I will consider two ways in which the classical theory of Bṛhaspati is modified in later Cārvāka, precisely in response to this worry. It is a worry which was present in the minds of the classical thinkers themselves.

Let me bring this phase in the discussion to a close by returning to Colebrooke and the lecture on Indian materialism he gave in London in 1827. It is striking now how many of the ideas that were to find a place in British emergentism are already there. The first of the British emergentists, J. S. Mill, used the example of chemical change to illustrate his idea of a 'heteropathic law' in *A System of Logic* (Mill 1843). Mill goes on to say that 'All organised bodies are composed of parts, similar to those composing inorganic nature, and which have even themselves existed in an inorganic state; but *the phenomena of life, which result from the juxtaposition of those parts in a certain manner*, bear no analogy to any of the effects which would be produced by the action of the component substances *considered as mere physical agents*' (1843, Bk. III, Ch. 6, §1; my italics). It seems likely that Mill, a person whose duties as a senior official of the East India Company included correspondence with Colebrooke, and who belonged with him to a circle of London literati based around the Royal Society, would have heard Colebrooke's lecture or read it when it was published in 1837, the very period he was working on *A System of Logic*. Colebrooke's work enjoyed an extremely wide circulation—even Hegel had some of his writings, and his translations of Sanskrit mathematical treatises were very well known to De Morgan and Boole. I cannot help but wonder if Bṛhaspati did not have a hand in the emergence of British emergentism.[19]

5. Transformation

A central element in mature Indian emergentism is the notion of a 'transformation' (*pariṇāma*). In many later sources the materialist is represented as holding the view that there is a transformation of those elements which are in the combination making up the body. Not mentioned in the earliest statements, it is appropriate to regard it as a development; and we can now begin to see what the motivations for such a development are. It is here that we should look for a positive characterization of the emergence relation.

Emergentism begins with the idea that systems which achieve appropriate levels of organizational complexity instantiate causal properties which are not exhibited by the components, whether as individuals or in aggregate. It is expressly stipulated that no familiar compositional model will render intelligible the emergence of these new properties. They are not scalar sums, as the mass of a whole is the scalar sum of the masses of its component parts, nor are they not vector sums, as the sum of a collection of forces results in a single new force. Nor are they mixtures, as the mixture of the colours of the parts results in the colour of the whole. The capacity to think is different in kind from any of the capacities or properties of the four elements, no matter how they are combined and synthesized. This is why Mill speaks of *heteropathic* laws and Broad of *trans-ordinal* laws; of which Bṛhaspati's 'thinking is from the elements' is alleged to be an instance. The emergentists' much-favoured example is chemical synthesis: for example, the emergence of salt and water from a reaction involving two quite different compounds, or, as our author said, the emergence of alcohol's powers from a process whose ingredients are sugar, yeast, barley, and water.

Searle argues that micro-neuronal features are *causally* sufficient for the instantiation of macro-mental features, and that this is what it is for the mental to 'causally supervene' on the physical. Mental properties are 'system features [which] cannot be figured out just from the composition of the elements and environmental relations; they have to be explained in terms of the causal interactions among the elements' (Searle 1992, p. 126). Searle claims that this is enough to describe them as emergent properties of the system, but distinguishes *his* concept of emergence (which he calls 'causal

emergence') from what he describes as a 'more adventurous' conception, according to which an emergent feature such as consciousness could cause things that could not be explained by the causal behaviour or the neurons: 'The naive idea is that consciousness gets squirted out by the behaviour of the neurons in the brain, but once it has been squirted out, it then has a life of its own.' The difficulty with Searle's account is that *neither* of his two conceptions of emergence is adequate: 'causal emergence' is too weak a notion, failing to sustain a robust explanation of the autonomy of emergent features, while the 'more adventurous' conception fails to do justice to the requirement that emergent features are dependent on the micro structures from which they emerge.[20] Relatedly, the idea that the relation between mental and physical properties is one of 'material causation' (*upādāna-kāraṇa*) is not held, by later Indian materialists, to suffice for a satisfactory characterization of emergence. They recommend a conception of emergence distinct from either of the two distinguished by Searle.

It is a deeply held intuition that nothing completely new can come into existence—nothing can come into existence which cannot be understood in terms of the nature of fundamental components and the ways they can be combined. What had formerly seemed mysterious about chemical reaction no longer surprises us, with our much better understanding of the nature of chemical bonds and the structure of atomic matter. The Cārvāka hypothesis about transformation can be seen as a way to reconcile this attachment to homopathic law with the key features of emergentism. Without a transformation in the micro-base, a homopathic theory of the emergence of psychological capacities is driven inevitably in the direction of panpsychism, for (so the thought goes) a complex could not think if the elements do not, any more than a whole could have a mass if all its parts were massless. The panpsychist alternative to emergentism has indeed been taken seriously by a number of philosophers in recent times,[21] but our sources provide two strong counter-arguments. One is that any object at all should then have psychological capacities, and we are lacking a clear criterion why only some do and others do not. The other is that, within a single body, there will be many sites of awareness, but no 'governing principle' orchestrating them:

> Even as the power to intoxicate resides to a small measure in each part of the intoxicating liquor, so too [the materialist must claim that] thinking is to a small measure in the parts of the body. And then many things will be thinking in one body. But it is impossible for the respective aims of many thinking entities to act in conformity, any more than many flying birds, bound by a single cord but disposed to move in conflicting directions, are able to cross even the distance of a span, even though the capacity is there for them to do so. So too the body would be unable to do anything. (Vācaspati 1980, p. 767, 21–24; cf. 1996, p. 531, 13–19)

It is in order to provide a non-panpsychist but not epiphenomenalist explanation of mental causation that the transformation theory is introduced. Let us suppose that the blending or combining of the elements 'transforms' them in such a way that in their transformed state their combination, according to homopathic principles, instantiates psychological properties. Then it will be true to say that mental properties are reducible to the properties of the transformed physical base but equally true that they are irreducible to the properties of the untransformed base.

One of our sources says that the view is that 'matter, although insentient in its inert state, will be bestowed with consciousness when in a body transformed' (Jayanta 1982, pp. 201, 26–202,1). Another says that it is the view that thinking, although not observed in the material earth out there, is present in the elements as transformed in the form of a body (Śaṅkara 1917, p. 765, 7–8). As so expressed, the idea seems to be that the elements themselves acquire new causal powers when they are in a certain state, namely the state of jointly composing a body, powers that they did not have beforehand when they were in other combinations with other elements. This is different from the view that the body as a whole has powers which none of its parts have individually. It is instead the view that the parts themselves have new powers *conditionally* upon their membership of the whole.

There is a resonance of this idea in the way Galen distinguishes between resultant and emergent properties. He says:

> Consider the first elements. Even though these substrata are unable to perceive, a body capable of perceiving can at some point come into being, because they are able to act on each other and be affected in various ways in many successive alterations. For anything constituted out of many things will be the same sort of thing the constituents happen to be, should they continue

> to be such throughout; it will not acquire any novel characteristic from outside, one that did not also belong to the constituents. *But if the constituents were altered, transformed, and changed in manifold ways, something of a different type could belong to the composite that did not belong to the elements* . . . Consequently, something heterogenous cannot come from elements that do not change their qualities. But it is possible from ones that do . . . Therefore, it is necessary that that which is going to sense be constituted either (i) from first elements capable of sensation or (ii) from ones incapable of sensation, but naturally such as to change and alter. (Galen, *On the Elements according to Hippocrates*, 1.3, 70.15–25, 72.19–21, 74.14–17, trans. Caston, 1997, pp. 355–7; my italics)[22]

Here Galen distinguishes two possibilities. One possibility is that the power to sense is an additive, resultant property, a possibility which leads directly to panpsychism. The other possibility is that the power to sense is an emergent property, and Galen's commitment to the principle that 'something heterogenous cannot come from elements that do not change their qualities' leads to the conclusion that the elements must be transformed.

The early British emergentists also use the word 'transformation.' but seem to mean something rather different by it. Thus Samuel Alexander:

> physiological complexes of a sufficient complexity carry mind or consciousness. They may be said to be 'transformed' in the consciousness they carry . . . the parts are used up to produce something different from them and transcending them, but, used up as they are, they are not altered or superseded but subserve. In this special sense there is a 'transformation' of the parts in building up a higher existence, *but the parts remain what they were*. (Alexander 1920, p. 370; my italics)

Alexander clearly asserts that in his view the 'parts remain what they were.' Carl Gillett (2006) has proposed that one reads this as the claim that an emergent property partly determines which causal powers are *contributed* by the base properties, that the base properties contribute causal powers in a way that is conditional upon the fact that they realize an emergent property. What distinguishes the Indian transformation theory from Samuel Alexander's is its claim that the emergent property determines not only what causal powers the base properties 'contribute,' but what causal powers they actually possess. The idea is that the parts have new powers in virtue of being parts of the whole and therefore intelligible only in reference to the whole to which they have come to belong. What powers an element has is conditional on what combination it is in. Emergence by transformation is the idea that the elements have cognitive powers only when in the frame of a living body, powers they do not have in other sorts of combination or in no combination at all.

What this brings into view is the availability of a conception of emergence distinct from either of the two conceptions distinguished by Searle. The proposal motivating the transformation theory is that, when micro-entities come together in appropriately complex systems of organization, the micro-properties they instantiate are transformed so as to give rise to novel causal powers in the macro-entity they constitute. The emergence of conscious states is not merely a fact about our inability to predict the behaviour of very complex systems, nor is consciousness something which is just 'squirted out.' It is a fact about the powers of micro-entities when they belong to macrophysical structures.

REFERENCES

Alexander, Samuel. *Space, Time, and Deity: the Gifford Lectures at Glasgow, 1916–1918* (London: Macmillan, 1920), 2 vols.

al-Faḍl, Abu, "Aīn-i Akbarī," in *The Aīn-i Akbarī*, H. Blochmann, ed., (Calcutta: Asiatic Society of Bengal, 1873–1907), 3 vols.

Beckermann, Ansgar, Hans Flohr, and Jaegwon Kim, eds., *Emergence or Reduction?* (Berlin and New York: de Gruyter, 1992).

Bedau, Mark, "Weak Emergence," in *Philosophical Perspectives*, vol. 11: "Mind, Causation and World," Oxford: Blackwell, 1997. pp. 375–99.

_____ "Downward Causation and Autonomy in Weak Emergence," in *Principia Revista Internacional de Epistemologica*, 6; reprint in *Emergence: Contemporary Readings in Philosophy and Science*, Mark Bedau and Paul Humphreys, eds., (Cambridge, MA: MIT Press, 2008), pp. 155–88.

Bedau, Mark, and Paul Humphreys, eds. *Emergence: Contemporary Readings in Philosophy and Science* (Cambridge, MA: MIT Press, 2008).

Bhattacharya, Ramkrishna, "Cārvāka Fragments: A New Collection," *Journal of Indian Philosophy* 30.6 (2002): pp. 597–640.

Broad, C. D. *The Mind and its Place in Nature* (London: Routeledge & Kegan Paul, 1925).

Burnyeat, Myles, "Is an Aristotelian Philosophy of Mind Still Credible?" in Nussbaum and Rorty 1992, pp. 15–26.

Cakradhara, "Granthibhanga," in *Nyāyamañjarī with the Commentary "Granthibhaṅga" by Cakradhara*, Gaurinath Shastri, ed., (Varanasi: Sampurnananda Sanskrit University, 1982), 3 vols.

Cardona, George, "*Anvaya* and *vyatireka* in Indian Grammar," *The Adyar Library Bulletin* 31–2 (1967–68): pp. 313–52.

_____ "On Reasoning from *anvaya* and *vyatireka* in Early Advaita," in Malvania and Shah 1981, pp. 79–104.

Carrier, Martin, Gerald J. Massey, and Laura Ruetsche, eds., *Science at Century's End* (Pittsburgh: University of Pittsburgh Press, 2000).

Caston, Victor, "Aristotle and Supervenience," *Southern Journal of Philosophy* 31 (1992): supplement, pp. 107–35.

_____ "Epiphenomenalisms, Ancient and Modern," *Philosophical Review* 106.3 (1997): pp. 309–63.

_____ "Dicaearchus' Philosophy of Mind," in Fortenbaugh and Schütrumpf 2001, pp. 175–93.

Chalmers, David. *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford University Press, 1996).

_____ "Strong and Weak Emergence," in Clayton and Davies 2006, pp. 244–54.

Cicero, Marcus Tullius. *Tusculan Disputations*, J. E. King, trans., (Cambridge, MA: Loeb Classical Library, 141. Harvard University Press, 1927).

Clayton, Philip, and Paul Davies, eds. *The Re-Emergence of Emergence* (Oxford: Oxford University Press, 2006).

Colebrooke, H. T. *Miscellaneous Essays* (London: W. H. Allen, 1837), 2 vols.

Cowell, E. B., "The Cārvāka System of Philosophy," *Journal of the Asiatic Society of Bengal* 31 (1862): pp. 317–90.

Crane, Timothy, "The Significance of Emergence," in Gillett and Loewer 1999, pp. 207–24.

Davidson, Donald. *Essays on Actions and Events* (Oxford: Clarendon Press, 1980).

Davies, Paul, "The Physics of Downward Causation," in Clayton and Davies 2006, pp. 35–51.

Dretske, Fred. *Explaining Behaviour: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988).

Fortenbaugh, William W., and Echart Schütrumpf, eds., *Dicaearchus of Messana: Text, Translation, and Discussion* (New Brunswick, NJ: Transaction Publishers, 2001).

Galen, *On the Elements According to Hippocrates*, Phillip De Lacy, trans., (Berlin: Akademie Verlag, 1996).

Gautama, "Nyāya-sūtra," in *Gautamīya-nyāya-darśana with Bhāṣya of Vātsyāyana*, Anantalal Thakur, ed., (Delhi: Indian Council of Philosophical Research, 1997).

Gillett, Carl, "Samuel Alexander's Emergentism: Or, Higher Causation for Physicalists," *Synthese* 153.2 (2006): pp. 261–96.

Gillett, Carl, and Barry Loewer, eds., *Physicalism and Its Discontents* (Cambridge: Cambridge University Press, 1999).

Humphreys, Paul, "Aspects of Emergence," *Philosophical Topics* 24.1 (1996): pp. 53–70.

_____ "How Properties Emerge," *Philosophy of Science* 64 (1997a): pp. 1–17.

_____ "Emergence, Not Supervenience," *Philosophy of Science* 64 (1997b): pp. 337–45.

_____ "Extending Ourselves," in Carrier, Massey, and Ruetsche 2000, pp. 13–32.

Jayanta, "Nyāyamañjarī," in *Nyāyamañjarī with the Commentary "Granthibhaṅga" by Cakradhara*, Gaurinath Shastri, ed., (Varanasi: Sampurnananda Sanskrit University, 1982), 3 vols.

Kamalaśīla. *Tattva-saṅgraha-pañjikā*, Swami Dwarikadas Shastri, ed., (Varanasi: Bauddhabharati, 1968).

Kim, Jaegwon. *Mind in a Physical World: An Essay on the Mind-body Problem and Mental Causation* (Cambridge, MA: MIT Press, 1998).

_____ "Making Sense of Emergence," *Philosophical Studies* 95 (1999): pp. 3–36; reprint in *Emergence: Contemporary Readings in Philosophy and Science*, pp. 127–54.

_____ "Emergence: Core Ideas and Issues," *Synthese* 151.3 (2006a): pp. 547–59.

_____ "Being Realistic about Emergence," in Clayton and Davies 2006, pp. 189–202.

Lowe, E. J. *Subjects of Experience* (Cambridge: Cambridge University Press, 1996).

_____ *Personal Agency: The Metaphysics of Mind and Action* (Oxford: Oxford University Press, 2008).

Macdonald, Cynthia, and Graham Macdonald, "Emergence and Downward Causation," in Macdonald and Macdonald 2010b, pp. 139–68.

Macdonald, Cynthia and Graham Macdonald, eds. *Emergence in Mind* (Oxford: Oxford University Press, 2010b).

McLaughlin, Brian, "The Rise and Fall of British Emergentism," in *Emergence: Contemporary Readings in Philosophy and Science*, pp. 19–60. Originally published in Beckerman, Flohr, and Kim 1992.

_____ "Emergence and Supervenience," in *Emergence: Contemporary Readings in Philosophy and Science*, pp. 81–98. Originally published in *Intellectica* 25 (1997).

Malvania, D., and N. J. Shah, eds., *Studies in Indian Philosophy* (Ahmedabad: L. D. Institute of Indology, 1981).

Mill, John Stuart. *A System of Logic, Ratiocinative and Inductive* (London: John W. Parker, 1843) 2 vols.

Morgan, C. Lloyd. *Emergent Evolution* (London: Williams and Norgate, 1923).

Muir, John, "Verses from the *Sarva-darśana-sangraha*, the *Vishnu Purāna*, and the *Rāmāyana*,

illustrating the tenets of the Chārvākas, or Indian Materialists, with some Remarks on Freedom of Speculation in Ancient India," *The Journal of the Royal Asiatic Society of Great Britain and Ireland* 19 (1862): pp. 299–314. (Read to the society on Dec. 14, 1861.)

Nagel, Thomas, *Mortal Questions* (Cambridge: Cambridge University Press, 1979).

Noordhof, Paul, "Emergent Causation and Property Causation," in Macdonald and Macdonald 2010b, pp. 69–99.

Nussbaum, Martha, and Amélie Rorty, eds., *Essays on Aristotle's De Anima* (Oxford: Clarendon Press, 1992).

O'Connor, Timothy, "Emergent Properties," *American Philosophical Quarterly* 31 (1994): pp. 91–104.

_____ *Persons and Causes* (Oxford: Oxford University Press, 2000).

O'Connor, Timothy, and John Ross Churchill, "Is Non-reductive Physicalism Viable within a Causal Powers Metaphysic?" in Macdonald and Macdonald 2010b, pp. 43–60.

O'Connor, Timothy, and Hong Yu Wong, "The Metaphysics of Emergence," *Noûs* 39 (2005): pp. 658–78.

Plutarch, "Against Colotes," in *Moralia*, vol. xiv, B. Einarson, trans., (Cambridge, MA: Loeb Classical Library, 428 Harvard University Press, 1967).

Prabhācandra. *Prameya-kamala-mārtaṇḍa*, Nyaya Shastri Mahendrakumar, ed., (Delhi: Sri Satguru Publications, 1990), 3rd ed.

_____ *Nyāya-kumuda-candra*, Nyaya Shastri Mahendrakumar, ed., (Delhi: Sri Satguru Publications, 1991), 2 vols. 2nd ed.

Rāmatīrtha, "Vidvamanorañjanī," in *The Vedānta-sāra of Sadānanda together with the commentaries of Nṛsiṃhā-sarasvatī and Rāmatīrtha*, G. A. Jacob, ed., (Bombay: Tukaram Jayaji, 1911).

Reydams-Schils, Gretchen, ed., *Plato's Timaeus as Cultural Icon* (Indiana: Notre Dame Press, 2003).

Rueger, Alexander, "Physical Emergence, Diachronic and Synchronic," *Synthese* 124 (2000a): pp. 297–322.

_____ "Robust Supervenience and Emergence," *Philosophy of Science* 67 (2000b): pp. 466–89.

Śaṅkara, "Brahma-sūtra-bhāṣya," in *The Brahma-sūtra-bhāṣya with the Commentaries Bhāmatī, Kalpatarū and Parīmāla*, N. A. K. Sastri and V. L. Sastri Pansikar, eds., (Bombay: Nirmaya Sagar Press, 1917).

Searle, John. *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992).

Sextus Empiricus, "Adversus Mathematicos VII," in *Sextus Empiricus II*, R. G. Bury, trans., (Cambridge, MA: Loeb Classical Library, 291: Harvard University Press, 1935).

Shoemaker, Syndey, "Kim on Emergence," *Philosophical Studies* 108 (2002): pp. 53–63.

_____ *Physical Realization* (Oxford: Oxford University Press, 2007).

Silberstein, Michael, "Emergence and the Mind-body Problem," *Journal of Consciousness Studies* 5.4 (1998): pp. 464–82.

_____ "In Defence of Ontological Emergence and Mental Causation," in Clayton and Davies 2006, pp. 203–26.

Silberstein, Michael, and John McGeever, "The Search for Ontological Emergence," *The Philosophical Quarterly* 49 (1999): pp. 182–200.

Sorabji, Richard, "The Mind-body Relation in the Wake of Plato's *Timaeus*," in Reydams-Schils 2003, pp. 152–62.

_____ *The Philosophy of the Commentators, 200–600 AD* (London: Duckworth, 2005), 3 vols.

_____ *Philoponus and the Rejection of Aristotelian Science* (London: Institute of Classical Studies, 2010), rev. 2nd ed.

Strawson, Galen, "Being Realistic: Why Physicalism Entails Panpsychism," *Journal of Consciousness Studies* 13.10–11 (2006): pp. 3–31.

Strawson, Peter F. *Individuals* (New York: Anchor Books, 1963).

Thompson, Evan. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (Cambridge, MA: Belknap Press, 2007).

Uddyotakara. *Nyāya-vārttika*, Anantalal Thakur, ed., (Delhi: Indian Council of Philosophical Research, 1997).

Vācaspati, Miśra, "Bhāmatī," in *Brahma-sūtra-śāṅkara-bhāṣyam*, Jagdish Lal Shastri, ed., (Delhi: Motilal Banarsidass, 1980).

_____ *Nyāya-vārttika-tātparyaṭīkā*, Anantalal Thakur, ed., (Delhi: Indian Council of Philosophical Research, 1996).

Van Cleve, James, "Mind-Dust or Magic? Panpsychism Versus Emergence," *Philosophical Perspectives*, vol. 4: 'Action Theory and Philosophy of Mind.' pp. 215–26. 1990

Varela, Francisco, Evan Thompson, and Eleanor Rosch. *Embodied Mind: Cognitive Science and Human Experience* (Cambridge, MA: MIT Press, 1991).

Wilson, Jessica, "Supervenience-based Formulations of Physicalism," *Noûs* 39.3 (2005): pp. 426–59.

Wong, Hong Yu, "Emergents from Fusion," *Philosophy of Science* 73 (2006): pp. 345–67.

APPENDIX

Authors and their dates

Dicaearchus (350–285 BCE) Aristotelian epiphenomenalist

Aristoxenus (fl. 335 BCE) Aristotelian epiphenomenalist

Galen (129–216 CE) Physician/Philosopher

Alexander (c. 200 CE) Aristotelian commentator

Bṛhaspati (date unknown[32]) Indian Materialist

Śaṅkara (c. 710) Vedāntin

Kamalaśīla (c. 740–795) Buddhist
Bhaṭṭa Udbhaṭa (c. 800) Materialist/Nyāya
Jayanta (c. 870) Nyāya
Vācaspati (c. 960) Polymath/Nyāya
Prabhācandra (980–1065) Jaina
H. T. Colebrooke (1765–1837); J. S. Mill (1806–1873); G. H. Lewes (1817–1878); C. Lloyd Morgan (1852–1936); Samuel Alexander (1859–1938); C. D. Broad (1887–1971)

NOTES

1. J. S. Mill 1854, and then Lewes 1875, Morgan 1923, Alexander 1920, and Broad 1925.
2. See for example Van Cleve 1990; O'Connor 1994; McLaughlin 1997; Kim 1999, 2006a, 2006b; Chalmers 1996, 2006; Crane 1999; Shoemaker 2007; Macdonald and Macdonald 2010a. Humphreys 1997, O'Connor 2000, and Lowe 1996, however, defend non-supervenience-based accounts.
3. I will discuss an alleged conception of emergence which denies this, so-called 'weak emergence.' below.
4. As Mark Bedau observes, 'emergent phenomena are Janus-faced; they depend on more basic phenomena and yet are autonomous from that base' (Introduction to Bedau and Humphreys 2008, p. 6); he elsewhere identifies dependence and autonomy as the two hallmarks of emergence (2003, p. 155).
5. Pronounced Bri-hus-puti; date unknown (but see Appendix).
6. athātas tattvaṃ vyākhyāsyāmaḥ । pṛthivyāpastej ovāyurititattvāni । tatsamudāyeśarīrendriyaviṣay asaṃjñaḥ । tebhyaś caitanyam । kiṇvādibhyo madaśaktivat । caitanyaviśiṣṭaḥ kāyaḥ puruṣaḥ । śarīrād eva । śarīre bhāvāt । (Bhattacharya 2002, pp. 603–4).
7. I take it that emergence as a philosophical thesis about the nature of metaphysical dependence on the physical is independent of the truth of any particular physical theory. For this reason I reject the suggestion that ancient philosophies of mind are no longer credible because ancient physics is not (Burnyeat 1992). Burnyeat's argument, in any case, is based on specific details of Aristotle's account of the physiology of vision, which, he alleged, prevent us from finding in him a proto-functionalist analysis of perception.
8. janmavaicitryabhedāj jagad api vicitram । mayūracandrakavat (Bhattacharya 2002, p. 604).
9. Bṛhaspati reserves for the term *puruṣa* 'human being' the same sense that P. F. Strawson does the term 'person,' that is with reference to specimens of a type of entity 'such that to each individual of that type there must be ascribed, or ascribable, *both* states of consciousness *and* corporeal characteristics' (Strawson 1963, p. 104).
10. Aristoxenus, another supporter of the *harmonia* theory, and someone who went with Alexander the Great to India, took the tuning to be of the organs and limbs rather than of the four elements within the body.
11. Caston 2001, p. 185: 'He accepts Aristotle's claim that a *harmonia* cannot have causal powers. But he does not think that this is a reason to reject the *harmonia* theory; if anything, it is a reason to change our views about the soul. He thinks that while there are mental events, they are completely inefficacious—their alleged effects are to be accounted for solely in terms of the powers of the body. Dicaearchus' position is that of the modern epiphenomenalist.' Caston mentions a passage from Plutarch's *On Desire and Grief* which attributes the view, apparently to Dicaearchus, that 'Some straightforwardly extend belief and calculation into the body, saying that the soul is not a cause at all, but that it is rather by the difference, quality, and power of the body that such things come about' (Caston 2001, p. 185; cf. Caston 1997, p. 345).
12. I owe this information to Richard Sorabji. See further his 2003; 2005, vol. 1, pp. 194, 201, 202; 2010, pp. 33–4.
13. Gautama 1997, p. 203, 3 (*Nyāya-sūtra* 3.2.46). Compare Śaṅkara 1917 (*Brahma-sūtra-bhāṣya* 3.3.53).
14. In the modal rather than the possible worlds formulation, given two families of properties A and B, A *weakly* supervenes on B if and only if necessarily, if anything *x* has some property F in A, then there is at least one property G in B such that *x* has G, and everything that has G has F; and A *strongly* supervenes on B if and only if necessarily, if anything *x* has some property F in A, then there is at least one property G in B such that *x* has G, and *necessarily* everything that has G has F.
15. Compare with Kim's formulation of what he calls the Principle of Downward Causation: 'To cause any property (except those at the very bottom level) to be instantiated, you *must* cause the basal conditions from which it arises' (1999, p. 24; my italics).
16. The locative absolute in Sanskrit, like the genitive absolute of Greek, can have a conditional, causal, temporal, or circumstantial force. Caston has observed that the genitive absolute is used, in ancient formulations of supervenience, with conditional force, expressing an antecedent (1997, p. 335). Here the locative absolute is being used in the same way. Caston has also pointed out that 'Aristotle . . . might have made his claim with the outermost necessity operator left implicit; philosophers often overlook this operator when speaking more loosely' (1992). Something very similar occurs here. An agreement in presence and absence is elsewhere described as a 'rule' or a 'principle' (*niyama*) (Uddyotakara says, for instance, that 'material things possessing weight fall because of it: this is [a case of] a rule'). It follows that the supervenience relation in emergence is here affirmed to carry nomologically rather than modally strong necessity (agreeing thereby with Noordhof 2010, pp. 71–2, that supervenience in emergence is *nomologically* strong).
17. See Bedau 1997, 2003; cf. Rueger 2000a, 2000b.
18. Chalmers (2006, pp. 252–3) considers a definition of weak emergence based on complexity ('Weak emergence is the phenomenon wherein complex, interesting high-level function is produced as a result of combining simple low-level mechanisms in simple ways') but prefers a more overtly epistemological definition, resting on notions of *interest* and *unexpectedness* ('A weakly emergent property

of a system is an interesting property that is unexpected, given the underlying principles governing the system'). He recommends that strong emergence is best characterized as 'non-deducibility even in principle,' and suggests that consciousness is the only strongly emergent phenomenon, all other examples of emergence being weak. The conception of emergence I am developing will be stronger than Chalmers's 'weak emergence,' but weaker than his 'strong emergence.'

19. Colebrooke's primary source for Cārvāka, Rāmatīrtha's commentary on the *Vedānta-sāra*, was first published in 1828. It was translated into English by Ram Mohun Roy in 1832 and into German by Othmar Frank in 1835. Two influential British Indologists, J. R. Ballantyne and A. E. Gough, published translations in subsequent decades. Thus, classical Indian emergentism was readily available to English-speaking audiences in the early nineteenth century.
20. Causal emergence has also been recommended in O'Connor and Wong 2005; Wong 2006.
21. Nagel 1979, pp. 181–95; G. Strawson 2006; Van Cleve 1990. Nagel's argument for panpsychism goes as follows: Human beings are complex systems composed entirely of matter [Materialism, Anti-Dualism]. Mental properties are not logically implied by any physical properties [Anti-Reductivism]. Human beings do have mental properties [Anti-Eliminativism]. There are no emergent properties [Anti-Emergence]. *Therefore*, the basic physical constituents of the universe have mental properties [Panpsychism].
22. Although rightly seeing in the passage an early anticipation of the distinction between emergent and resultant properties, Caston curiously does not remark on the role to which Galen accords the concept of transformation. Neither does Kim, who quotes the passage in his 2006b, but glosses it in such a way that the idea of transformation entirely disappears: 'Galen is saying that a composite object made up of simpler constituents, when these constituents enter into special complex relationships ("act on each other and be affected in various ways"), can come to exhibit a novel property ("something of a different type") not possessed by its constituents' (2006b, p. 189). It seems to me that Caston and Kim are too keen to read Galen as anticipating *modern* understandings of emergence, and in doing so fail to notice an idea which an examination of the Indian theory makes vivid.
32. The first known reference to Bṛhaspati is from the sixth century. He composed his text in the sūtra style, and the usual period of production for texts in that style is thought to be the interval between 200 BCE and 200 CE. Such texts were often distillations of ideas already in circulation. There are formal resemblances between Bṛhaspati's text and other sūtra texts whose dates have been more precisely ascertained; in the *Nyāyasūtra*, for example, one finds the sentence 'The senses are from the material elements (*bhūtebhyaḥ*).' It is reasonable to speculate, therefore, that Bṛhaspati is no later than 200 CE. For further speculation, see Muir 1862, Cowell 1862.

21 Post-Physicalism

Barbara Montero

Introduction

What is the problem, inherited from Descartes, that we now call 'the mind-body problem'? In his most recent book, Jaegwon Kim provides an answer with which many would agree. 'Through the 70s and 80s and down to this day,' Kim tells us, 'the mind-body problem—our mind-body problem—has been that of finding a place for the mind in a world that is fundamentally physical' (Kim 1998, p. 2). This problem, which at one time was at home mainly in departments of philosophy, is now studied by a broad range of disciplines. One finds, for example, neuroscientists arguing that certain discoveries about the brain show that consciousness is physical; researchers in artificial intelligence claiming that because human thought can simulated by complex computers, thought requires nothing beyond the physical; and evolutionary biologists declaring that insights into the evolution of the mind indicate that it must be fundamentally physical. But what does it mean to be physical? While the basic results of the research being done may be clear enough, how are we to interpret the further claim 'and this shows that the mind is physical'? The answer is that we have no idea.

I am going to argue that it is time to come to terms with the difficulty of understanding what it means to be physical and start thinking about the mind-body problem from a new

From *Journal of Consciousness Studies*, 8:2, pp. 61–80, 2001. Reprinted with permission of the author and the publisher.

perspective. Instead of construing it as the problem of finding a place for mentality in a fundamentally physical world, we should think of it as the problem of finding a place for mentality in a fundamentally nonmental world, a world that is at its most fundamental level entirely nonmental. The mind-body problem, I want to argue, is the problem of determining whether mentality can be accounted for in terms of nonmental phenomena. In other words, it is the question, 'is mentality a fundamental feature of the world?'[1]

I: The Current State of the Debate

Currently most philosophers working on the mind-body problem see the debate in terms of the physical and the nonphysical: the question most are concerned with is whether mentality is fundamentally *physical*.[2] Indeed, since most think that the mind *must* be physical, the project they are engaged in is not so much arguing that the mind is physical, but, rather, trying to show how the mind could be physical (given that it is). And so, whether the account of mentality that physicalists propound is expressed in terms of reduction, realization, identity, supervenience, explanation or even elimination, the goal is to provide a plausible theory of mentality (or, as the case may be, a theory that accounts for what we mistakenly took to be mentality) that is compatible with the view that the world is fundamentally physical. For example, if one thinks that it is incumbent on physicalists to *explain* mentality then the explanation, it is thought, must make reference exclusively to physical phenomena; if one thinks supervenience suffices for physicalism, then the supervenience base must be entirely physical; and so forth. But what does it mean to be physical? It seems that those who take the central concern of the mind-body problem to be the relationship between mental properties and physical properties—and if Kim is right, this is just about every-one—should have at least a rough idea of what it means to be physical, not necessarily a strict definition, but at least a notion of the physical that excludes some, if not actual, then at least possible, phenomena from being physical. For if we cannot even conceive of something being nonphysical, it is difficult to grasp what physicalists could be arguing for—to say nothing of what that they could be arguing against.[3]

It is not at all clear, however, that physicalists can provide even this minimal condition. Current physics, which posits such things as particles with no determinate location, curved space-time, and wave-particle duality, tells us that the world is indeed more ghostly than any ghost in the machine. And if the existence of ghostly phenomena does not falsify physicalism it is difficult to say what would. As Richard Healey puts it, '[the] expanding catalogue of elementary particle states of an increasingly recondite nature seems to have made it increasingly hard for the physicists to run across evidence that would cast doubt on a thesis of contemporary physicalism stated in terms of it' (Healey 1979, p. 208). In other words, if such things as one-dimensional strings and massless particles are physical, it is difficult to say what wouldn't be. Bertrand Russell made this basic point back in 1927: 'matter,' he said, 'has become as ghostly as anything in a spiritualist's séance.'[4] And over the past seventy years Russell's point has, if anything, been reinforced. Presumably things could change. Philosophy, as we all know, is not noted for its rapid progress and perhaps in another seventy years or so we will have a clear idea of what it means to be physical. However, it seems to me that until such clarification comes about, we ought to rethink the project of accommodating the mental in the physical world. That is, we ought to rethink what Kim tells us is 'the shared project of the majority of those who have been working on the mind-body problem over the past few decades' (Kim 1998, p. 2).

Not surprisingly, most physicalists are of a somewhat different opinion. While many physicalists admit that our understanding of what it means to be physical is rather tenuous, they usually think that the notion, and thus the crux of the debate, is clear enough. The mind-body problem, according to most physicalists, is the problem of explaining how the mind can be physical, where what counts as physical is given to us by science. In John Searle's words, the mind-body problem is the problem of locating mentality 'within our overall "scientific" conception of the world.'[5] And so, it does not matter what kinds of ghostly and bizarre phenomena science may posit, for it is science itself that serves as a reality test. Searle thinks mentality passes the test because mentality, he argues, is 'as much part of our biological natural history as digestion' (Searle 1992). Others, however, are a bit harsher in their grading policy. According to Patricia Churchland,

for example, it is premature to say that every aspect of what we now think of as mentality can be accommodated in our scientific world-view (and for Churchland the relevant science here is neuroscience) since, for all we know, certain aspects of mentality might fail the test and go the way of phlogiston (Churchland 1995). Yet as different as their views may be, both Searle and Churchland, as well as most other physicalists, abide by Wilfred Sellars' well known dictum, 'in the dimension of describing and explaining the world, science is the measure of all things, of what is that it is, and of what is not that it is not.'[6] Physicalists may disagree about just how far to take this claim: must we be 'nothing butists,' or can we accept an ontology that goes beyond science as long as it is related to the posits of science 'in the proper way'?[7] However, when it comes to fundamental ontological matters, they are, for the most part, united: the ultimate authority is science.

But what is meant here by 'science'? Physicalists usually shy away from expressing their views about which specific theories will account for the fundamental nature of, as it were, everything. And this, of course, is the safest strategy. For as David Lewis advises, physicalists should 'side with physics, but not take sides within physics' (Lewis 1983, p. 364). Samuel Guttenplan advocates this strategy as well; in his words, 'all we [physicalists] are claiming is that any phenomenon that is a genuine happening in this world is in principle explicable by a science albeit by a science that might be quite different from any we now have at our disposal' (Guttenplan 1995, p. 77). But if this is all that physicalists are claiming, it is difficult to see what prevents *anything* from being physical: if physics (correctly) tells us that some things have no mass or no determinate spatial location, well then, physicalists will say, those things will still count as physical. Even if physics were to one day reveal that our current theory of space-time is mistaken and that space and time actually are distinct so that some phenomena have temporal, but not spatial properties, then physicalists, I assume, would say that those things too, if they actually exist, will be physical. Even more, if, as some physicists have begun to speculate, there is some sort of nonspatial, nontemporal stuff out of which space-time itself emerges, physicalists will once again declare victory.[8] But if this is so, it seems that the strategy of simply siding with science, whatever science may ultimately say, is so safe as to bestow physicalism with what Popper thought was the very unscientific virtue of being, even in principle, unfalsifiable. Perhaps the deep eternal truths that are the domain of philosophy as well as mathematics are not at all likely to be falsifiable. Yet it seems that without any restrictions on how the science in question is to progress, or on what entities and properties it is to incorporate, physicalism, that is, the view that everything is physical, becomes not only unfalsifiable, but also trivial.[9] That is, without any restrictions whatsoever, the view that everything is physical ends up as the view that everything exists. And this, it seems to me, is a position that most philosophers, save, of course, for Meinongians, are not interested in discussing.

While a number of physicalists, including Lewis himself, have tried to avoid this obstacle in their formulations of physicalism, I think that ultimately there is no way around it.[10] As long as one defines the physical in relation to what science tells us about the world, the problem of explaining what it means to be physical in the context of the mind-body problem, a problem I call 'the body problem,' currently has no solution.[11] But what is left of the mind-body problem if we have no notion of body? In other words, is there room for the mind-body problem in a post-physical world?[12]

II: Is There Still a Mind-Body Problem?

One might think that the only reasonable conclusion to draw from the view that we have no notion of the physical is that we should give up the mind-body problem altogether: declare it dissolved and move on to other, hopefully better-defined, problems. And as far as I know, most of those who argue that we have no philosophically useful notion of the physical are, it seems inevitably, drawn to this conclusion. Noam Chomsky is a good example. He tells us, 'we can speak intelligibly of *physical* phenomena (processes, etc.) as we speak of the *real* truth or the *real* world, but without supposing that there is some other truth or world' (Chomsky 1998, p. 438). And he takes this to mean, 'we have no coherent way to formulate issues related to the "mind-body problem"' (Chomsky 1995, p. 5; see also Chomsky 1993). Similarly, Bas van Fraassen argues that the fact that physicalists will usually count '*whatever science comes up with*' as physical shows that the thesis of physicalism lacks content (van

Fraassen 1996, p. 167). Chris Daly, who, in a recent paper, argues quite forcefully that we have no notion of a physical property, concludes, 'no debate between physicalism and dualism can even be set up' (Daly 1998, p. 213; also see Scheffler 1950). While Tim Crane and Hugh Mellor, after finding flaws with a wide variety of proposals for defining physicalism, conclude that their paper 'should really be the last paper on the subject' (Crane and Mellor 1990, p. 83). The pattern is clear. And it is not at all difficult to see the motivation behind it: for if we have no notion of the physical, there seems to be little use in asking how the mind could be physical and, thus, little point in discussing the mind-body problem. But is this the only conclusion one can draw? Must our inability to solve the body problem lead to the demise of the mind-body problem as well?

To be sure, one obvious worry about concluding that we must abandon the mind-body problem is that, as a matter of fact, very few will follow suit. Philosophy, it has been said, has a penchant for burying its undertakers, and despite repeated pronouncements of the death of the mind-body problem, most people feel that a problem of some sort—perhaps of a very deep sort—remains. Even Crane and Mellor realize that this creates some tension in their view. For after stating quite boldly that their paper should definitely be the last on the topic, they also sheepishly admit that they actually know it will not. And here they were certainly right. Since their paper came out, about ten years ago, the question, 'what is the fundamental nature of the mind?'—a question to which 'it is physical' is supposed to provide an answer—has, if anything, been even more widely discussed. But why is this, if, as the title of their paper proclaims, 'there is no question of physicalism'? Of course, the mere fact that many continue thinking about a problem does not show that a problem really exists. For it might be that no one has listened to Crane and Mellor's protests that we have no notion of the physical capable of grounding questions about whether the mind is physical. While there may be something to this, there is more to be said. For there actually *is* an interesting question to ask about the fundamental nature of the mind. It is just not the question of whether the fundamental nature of the mind is physical.

What other broad, philosophical questions can we ask about the fundamental nature of the mind—questions, that is, which could reasonably be thought to address the set of concerns that we have come to think of as the mind-body problem? Certainly, even if there were no philosophical problem called 'the mind-body problem,' there would still be specific questions about the mind left to investigate. For example, regardless of whether we have a notion of the physical, we may still arrive at a deeper understanding of our mental lives, perhaps by studying the relationship between consciousness and various neural process or, say, investigating which sorts of pains are correlated with A-fibre stimulation, which with C-fibre stimulation. Daly emphasizes this point towards the end of his paper. As he says, 'for even in absence of a principled account of the distinction between physical properties and all other properties, terms used to designate specific properties may be sufficiently well defined for us to raise specific issues [such as, how pain relates to C-fibre stimulation]' (Daly 1998, pp. 213–14). And I take it that many of those who take sides within science (rather than simply siding with science) are engaged in addressing specific questions—like those the neuroscientist asks—that do not depend on such a distinction. But are specific questions the only sorts of questions left? I think that they are not, for regardless of whether we know what it means to be physical, we can still ask whether mentality is a fundamental feature of the world. In other words, does mentality ultimately depend on nonmental phenomena or, as it were, is it mental all the way down?

As I see it, in its most general structure, this is the crux of the mind-body problem. Yet a glance through the literature, where one comes across numerous papers with titles such as, 'Can Science Explain Consciousness?' does not make this apparent (see, for example, Shear 1997). That is, the question of whether mentality is fundamental is crucial to the debate, yet rarely is it addressed directly. I think this is a mistake. And a serious one since physicalists aim to refute dualism, yet dualism is the view that mentality is fundamental.

To say that the question is rarely addressed directly, however, is not to say that philosophers are indifferent as to its outcome. For despite the fact that most recent discussions about the mind focus on whether mentality will be somehow subsumed under the scope of the scientifically acceptable, what I call 'the science question,' one often does find that the underlying concern in these debates is the question of whether mentality is a fundamental feature of the world (what *I* call 'the mind-body problem'). And, for the most part, philosophers' views on this

come down along party lines: the dualists are for it and the physicalists are against it.[13] For example, when Kim lays out the basic physicalist commitments, along with the claim that the mental supervenes on and is determined by the physical is the claim that there are 'no fundamental mental entities.'[14] David Chalmers also makes clear what side he is on: when addressing the question of whether it would be more accurate to call his view a version of physicalism, since he allows that the mental may in the future be accounted for by an expanded physics, he holds fast to the dualist classification because, as he says, his view admits 'phenomenal or protophenomenal properties as fundamental' (Chalmers 1996, p. 136; see also Foster 1989, pp. 1–15).

But, again, while many physicalists claim that mentality is not fundamental, few spend much time defending this claim. Rather, most focus on the science question, the question of whether science will account for the mind.[15] Yet these two views do not make the same cut: science may account for mentality (in as much as it accounts for any other fundamental feature of the world) but mentality may still be a fundamental feature of the world. Of course, physics, for the most part, does not posit anything fundamentally mental. However, it is not too much of a stretch of the imagination to see how it could. For example, if Wigner's hypothesis—the hypothesis that acts of pure consciousness (in other words, fundamentally mental entities) are required to explain the collapse of the wave function—were accepted, or, to put it more strongly, if it were true, there would be a sense in which consciousness fits perfectly into our scientific world-view: acts of pure consciousness would be just one of the many fundamental entities posited by physics.[16] Yet mentality would still be a fundamental feature of the world. Certain interpretations of the anthropic principle, a principle sometimes invoked to explain or at least constrain other explanations about why things are just as they are, also seem to take mentality as, at least explanatorily, fundamental. For example, it is sometimes claimed that the reason why a particular state of the carbon nucleus has the precise energy that it does is that if this value were only slightly greater or slightly less, human beings would never have developed and thus we would not be able to ask this very question. As such, the existence of human consciousness is taken as a starting point in explaining other aspects of the universe.[17] To be sure, both the anthropic principle and Wigner's hypothesis are highly controversial, and perhaps neither should be taken as part of physics. But I think they do illustrate a possibility: the possibility of how physics could incorporate mentality as a fundamental. And as long as physicalists accept this mere possibility, they are accepting the possibility that mentality could be accounted for by science, yet still be fundamental.[18] Arguing that science can in principle account for mentality, then, does not suffice to show that mentality is not fundamental; in this sense, an argument for physicalism is not an argument against dualism.[19]

But while physicalists, in claiming that science or physics will account for mentality, may assert a view that is not opposed to dualism, they nonetheless usually exclude the mental from their fundamental ontology. And since they do, they in fact are not simply deferring to science to tell us how things are. Rather, they are putting forth a substantive ontological thesis, namely, that mentality is not a fundamental feature of the world (regardless of what physics reveals). If physics posits things that cannot be analysed in terms of (or are not ultimately reducible to or determined by or constituted out of—pick your favourite dependence relationship) nonmental phenomena, physicalists typically will *not* go on to acknowledge those things as physical. Science may, indeed, be the measure of all things, but if science posits fundamental mental entities or properties, physicalists, I take it, throw in the towel. And so, despite much talk about the wonders science is capable of achieving, the crux of the mind-body problem is actually not the question 'is the mind physical?' (where this notion is tied to what science can achieve) but is rather the question, 'is the mind fundamentally nonmental?'[20]

III: Physicalism Without Naturalism = Fundamental Nonmentalism

Convincing physicalists that we need to focus on the question of whether mentality is fundamental may not be easy since physicalists are usually intent on trying to have things both ways. As I see it, physicalists have long struggled to find some middle ground between their desire to be naturalists, that is, their desire to defer to science for matters about fundamental ontology, and their desire to put forth a significant thesis about the mind, that is, their desire to express a view that at least amounts

to more than the claim that mentality fits into our scientific world-view, where our scientific world-view can turn out to encompass anything including, if it so happens, acts of pure consciousness. But it is very difficult to do the latter while upholding the former. In order to make their notion of physicalism a substantive claim about the world, some restriction needs to be placed on what counts as science.[21] Putnam makes this point forcefully: 'if no restraint at all is placed on what counts as a possible "fundamental magnitude" in future physics, then reference to soul or good could even be fundamental magnitudes in future physics!'[22] Yet placing *a priori* restrictions on science, on what it is and how it is allowed to progress seems blatantly anti-naturalistic: according to the deferential naturalist, what sorts of theories about the world, what sorts of entities, relations, and laws science can posit as well as what sorts methodologies it can avail itself of, will be determined by science itself, not by armchair philosophy. Given this conflict of interests, I say physicalists should abandon one or the other: naturalism or ontological significance.[23]

Naturalists might try to avoid this conflict by claiming that their intent is not to place restrictions on the posits of science but, rather, to make a prediction about its course, namely, that mentality will not show up as a fundamental.[24] But this consistency is purchased at a price. For to adopt a policy of strict noninterference and recede to mere prediction is to step out of the debate between physicalists and dualists. Some naturalists may not mind this retreat, yet I think that they are not the majority. Rather, many naturalists both positively assert that mentality is not fundamental and claim to defer all ontological matters to the scientist and, thus, struggle valiantly to remain faithful to an ontology that excludes fundamental mentality and to deferential naturalism. Yet this, I think, is something that cannot be done with consistency.

David Papineau is a good example of someone engaged in such a struggle. Papineau provides a clear account of which dependence relation he prefers: the mental, he claims, supervenes on and is token congruent with the physical. He takes supervenience to be a thesis about variation between intrinsic properties of systems: two systems cannot differ (across possible worlds that share our laws of physics) without differing in terms of their intrinsic physical properties. And he takes two properties to be token congruent if one realizes the other or if they are actually type identical.[25] But when it comes to explaining what he means by 'physical,' a conflict becomes apparent. His commitment to naturalism leads him to take a hands-off approach and to let the answer to this question come from within physics itself. Not today's physics, since he thinks that current physics is certainly inadequate, but rather a true and complete physics, a physics that he simply defines as 'the science of whatever categories are needed to give full explanations for all physical effects.'[26] But now the threat of triviality enters the picture: for if psychological categories are part of this science, Papineau's position loses its punch. That is, if psychological categories are part of the science that is needed to explain all physical effects, psychology will indeed supervene on the physical, but only because it will be part of the physical (there will be no change in psychological properties without a change in the physical base properties because the psychological, itself, will be a physical base property).

Papineau, however, well aware of this threat, tries to compromise. The mind, he claims, will be accounted for by science, yet the science will be one without psychological categories (Papineau 1993, p. 31). But it seems to me that this exclusion is really the whole game. The bottom line, it turns out, is not whether mentality can be accounted for by science. For when the notion of science is left entirely open-ended, as Papineau, being a good naturalist, is drawn to do, we can say nothing about whether psychological categories will be part of the *final* scientific dependence base. Rather, the bottom line is whether mentality can be accounted for without involving psychological categories themselves.

Robert Kirk's discussion of what he means by 'physical' exemplifies a similar conflict of interests. The physical, Kirk says, is simply 'whatever is posited by physics.' Yet, just to be safe, he also says, 'we can explicitly exclude all expressions that would ordinarily be counted as mental or psychological' (Kirk 1994, p. 78). But clearly one cannot leave everything up to the physicists while at the same time placing restrictions on what they can do. One can embrace naturalism wholeheartedly: take one's ontological commitments to reach only as far as what is sanctioned by science and thereby defer all substantial ontological questions. Or one can take a stance: reject naturalism and start being a little less deferential.

The middle ground that Papineau and Kirk try to set out, that is, leaving the job of making all substantial ontological hypotheses up to

the scientists *except* for the hypothesis that the mental is not fundamental, seems oddly *ad hoc*. Why should this bit of *a priori* reasoning be allowed and not others?[27] Some might say that it should be allowed because the hypothesis that mentality is fundamental is abhorrent to common sense, simply unimaginable. And perhaps this is so. For as Thomas Nagel has said, 'there is a deep-seated aversion in the modern "disenchanted" Weltanschauung to any ultimate principles that are not dead—that is, devoid of any reference to the possibility of life or consciousness' (Nagel 1996, p. 133). However, reasoning from what is or is not abhorrent to common sense is not usually a type of reasoning condoned by naturalists. According to the naturalist, scientific judgments are one thing and intuitions are something else. For example, naturalists may admit that it is intuitively difficult to understand how pain could be identical to, or even just constituted by some brain state. But this, they quickly point out, does not falsify physicalism. Newtonian gravity was difficult to imagine, naturalists often remind us, but this didn't stop Newton and likewise the fact that some hypothesis is unintuitive or difficult to imagine should not stop physicalism. But if naturalists reject reliance on what is or is not abhorrent to common sense in these situations, it seems that consistency should lead them to reject it in reasoning about whether mentality is a fundamental feature of the world. If the naturalist wants to leave everything up to science, then he should do so. For there is nothing wrong with adopting the strategy of 'let's just wait and see.' It is just that this strategy does not make for much of a debate.

It might seem, however, that the distinction between the naturalist who defers to the scientist and the armchair philosopher who denies the fundamental status of mentality is somewhat artificial. For isn't it the case that science, in fact, already tells us that mentality is not fundamental? To be sure, if we try to look up mentality in the 'Berkeley booklet,' the physicists' ever expanding little black book of the fundamental entities and properties known to date, we will find no listing.[28] But according to most physicalists, not being listed in today's Berkeley booklet does not exclude a fundamental entity or property from the physical realm. What matters is what will show up in the final edition. That is, physicalists do not merely claim that mentality is not *currently* classified by physicists as a fundamental feature of the world; rather, according to such philosophers as Papineau and Kirk, mentality will be *forever* unlisted in the physicists' little black book.[29] But it is difficult to see how naturalists can make such an assertion. The world as we know it is full of fundamental properties and fundamental mental properties (if there are such things) should, at least in principle, be no more (or for that matter, no less) mysterious, or necessarily outside the realm of science than any other fundamental properties.[30] Of course, there may be reasons to avoid such a conclusion. For example, considerations of simplicity may lead us to want explanations with as few primitive terms as possible and encourage us to do without mentality as a primitive term if possible. But the question of whether it is possible is the central point of the debate so we should not start out assuming this.

That said, one still might feel that there is something defeatist about the view that mentality is fundamental. For if one claims that mentality is fundamental isn't one, rather than presenting a possible solution to the mind-body problem, giving up on it?[31] I suppose this depends on what one means by 'giving up.' To be sure, fundamental principles or phenomena do, by their very nature, leave something unexplained. And in this sense, claiming that mentality is fundamental is tantamount to denying any possible further explanation of it, any explanation of it in terms of something else. Nevertheless, to successfully argue that mentality is fundamental is to provide a solution to the mind-body problem. That is, that answer to the question 'what is the relationship between mental phenomena and nonmental phenomena?' will be that the mental is fundamentally distinct from the nonmental.

Of course, the distinction between the mental and the nonmental may not be sharp. And in fact, physicalists may need to hold that it is not. Physicalists (of the noneliminative sort) think that the nonmental, arranged in the right way, as it were, gives us the mental; yet if you hold that the mental/nonmental distinction is sharp, it is very difficult to see how to bridge that divide. For example, if you are interested in explaining subjectivity and you take the line between the subjective and the objective (i.e., the nonsubjective) to be sharp, physicalism—or rather, fundamental nonmentalism—becomes very difficult, if not impossible to defend. This is true of the intrinsic/extrinsic divide as well: if it is sharp, and one takes the properties of consciousness to be intrinsic and physical properties to be extrinsic, it is very difficult to see how

an anti-physicalist view could fail to follow.[32] As a sharp distinction between the living and the non-living seems to lead us to posit a fundamental life-force, or *élan vital*, a sharp distinction between the mental and the nonmental seems to lead to dualism. So in debating the mind-body problem, we should focus on the mental/nonmental distinction but not presuppose that the distinction is sharp since the outcome of the debate may partially turn on this.

IV: Some Difficult Classifications

I have argued that physicalists are committed to the claim that mentality is not a fundamental feature of the world and that this commitment is incompatible with the naturalist's commitment to defer all substantial ontological questions to the physicists. The latter leads to a hands-off approach that basically amounts to stepping out of the debate, while the former presents in extremely general form a relatively clear way of posing the mind-body problem, a way that enables a *real* debate, that is, a debate that has more than one side. But two sides, while better than one, do not cover all possible positions and it may not be entirely clear how to classify some of them. For example where do we place Chalmers' suggestion that mentality may be accountable for in terms of something like 'proto-mentality,' and that 'proto-mentality' is a rock bottom feature of the world? For it apparently amounts neither to the view that mentality is fundamental nor to the view it is not fundamental. Similarly, where does one place the view that an explanation of the mental not only requires reference to lower level neurological processes but also requires reference to higher level features of the world, such as social features and natural selection? (See, for example, Wimsatt 1976; 1994; Clark 1997). While such positions illustrate borderline or hybrid cases, I do not think that the existence of such cases reduces the usefulness of setting out the two extremes.

But things can get tricky: What if one thinks that in order to account for mentality we need to invoke God? Proponents of this view certainly do not think of themselves as physicalists; yet it is not clear that they think of mentality as being fundamental either. Nevertheless, there is a sense in which they take mentality as fundamental feature since adverting to God often involves adverting to mentality. Perhaps for some believers, those who think of God as something like energy, or perhaps even the big-bang, this is not the case. If so, they fall on the other side of the debate.

Trickier still, however, is view that we need norms to account for mentality. Roughly speaking, this is the view that mentality—more specifically, intentionality—is fundamentally normative.[33] Thus normativity, and not necessarily mentality, is taken as fundamental. But, someone might object, the fundamentally normative is just as abhorrent to physicalists as the fundamentally mental. As with adverting to God, it is not clear on which side of the debate this view lies. But here, again, I would want to ask whether the account of mentality provided is an account of mentality in terms of the nonmental.[34] If intentionality is accounted for by norms that are themselves irreducibly intentional, this is an instance of taking intentionality to be irreducible. And if intentionality is accounted for by the nonintentional, but normative, this is simply an account of intentionality in nonintentional terms. To those who may protest that this latter position would nonetheless be antiphysicalistic, I can only ask them for their solution to the body problem. Without some understanding of what it means to be physical, these protests fall dead in their tracks. In any event, the possibility of a normative account of mentality provides no reason to return to the science question. For it is not clear that there is anything inherently unscientific about fundamental norms. If they exist, I suppose that they too will need to be included in that proverbial true and complete catalogue of the fundamental features of the world.

V: The Impact on the Debate

One might accept my formulation of the mind-body problem, but still wonder if it will have any interesting effect on the debate. If not, we might as well save the ink and leave everything as is. A look, however, at what some see as the most persuasive argument for physicalism indicates that we can't. Why believe in physicalism? One not uncommon answer is that the tremendous success and progress of the physical sciences gives us reason to think that physicalism is true. But the tremendous success of physics, while possibly very relevant to the outcome of the debate, does not settle the issue between physicalists and dualists. For it seems that physics could be tremendously successful in either case.

Moreover, I think that focusing on the mental/nonmental distinction rather than the physical/nonphysical distinction will also affect the debate about mental causation. The problem of mental causation is usually thought of as the problem of explaining how mental properties could be casually efficacious in a world that is fundamentally physical. The difficulty arises, it is thought, for anyone who thinks that (1) mentality exists but is not identical to anything physical, (2) there is no causal overdetermination, and (3) the physical world is causally closed, i.e. all physical effects (which have causes) have sufficient physical causes. Many philosophers have thought that there are good reasons to accept all three of these claims. But it seems to me that when we shift our focus to the mental/nonmental distinction, the reasons usually given for the third claim, the causal closure of the physical, no longer apply. For the reasons usually given for why we should believe in the causal closure of the physical are the reasons usually given for why we should believe in the causal closure of physics. As Kim puts it, if the physical world were not causally closed then 'to explain some physical events you must go outside the physical realm and appeal to nonphysical causal agents and laws governing their behavior!'[35] And what this entails, and I take it why Kim finds it exclamatory, is, as he says, 'complete physics would in principle be impossible, even as an idealized goal' (Kim 1996, p. 147). This is true when the physical is defined over physics, but it is not a reason to accept the causal closure of the fundamentally nonmental.[36]

Some might claim that one potentially unwelcome result of my formulation will be the demise of the identity theory—the old example being that pain is identical to C-fibre stimulation. For what could the identity theory amount to if we take the relevant distinction to be between the mental and the nonmental?[37] If we think of C-fibre stimulation, for example, as entirely nonmental, what could it mean to say that C-fibre stimulation is identical to something mental? Yet if you allow C-fibre stimulation to have an irreducible mental aspect, in what sense are you a physicalist? In general, isn't it hopelessly paradoxical to say that mental property *M* is identical to nonmental property *N*?[38] Perhaps it is, but my claim is not that we should replace the notion of *being physical* with the notion of *being nonmental*, from which it would follow immediately that the identity theory is impossible. Rather, it is that we should replace the notion of *being physical* with the notion of being *fundamentally* nonmental. As such, it seems at least possible to carve out a space for the identity theory: the identity theory would be true if both pain and C-fibre activity are, at least in some sense, mental yet the fundamental constituents of C-fibre stimulation (as well as pain) are entirely nonmental.

Most importantly, however, I think that focusing on the mental/nonmental distinction will facilitate an actual head-on debate between physicalists and dualists. As things stand, physicalists usually take themselves to be arguing against views about the mind that are anti-scientific, views that hold that mentality will forever be beyond the scope of science. Yet dualists often take themselves to be arguing against the view that mentality is not fundamental (regardless of whether it can be accounted for by science). So it is not surprising that the two sides of the debate often talk past each other. If we focus on the mental/nonmental distinction, this may change.

VI: Is the Mental/Nonmental Distinction Clearer Than the Physical/Nonphysical Distinction?

The change, I hope, would be for the better; yet it may not if the distinction between the mental and the nonmental is no clearer than the distinction between the physical and the nonphysical.[39] It is not an easy task to delineate the mental from the nonmental. However, I do think that this distinction is better off than the physical nonphysical distinction. Why do I think this? I could say, as one is prone to do when asked this question, that we are familiar with what is often classified as the qualitative aspect of mentality, that is, what it is like from our first-person perspective, for example, to feel pain, to see red, to taste chocolate, to have the unpleasant experience of being embarrassed or the wonderful experience of feeling proud, and so on. Yet to do so does little to convince those who resolutely deny having any understanding of phenomenal experiences that they actually know what they are like. And eliminativists are just such people: according to (the more radical) eliminativists, we couldn't have any understanding of mentality since there really isn't any such thing. And it is difficult to know what sort of argument one could give that would convince them to believe

otherwise (this is especially tricky when they claim that, strictly speaking, they have no beliefs).[40]

So let me try a new tack. An indication that we have a grasp of the mental is that while there may be no agreed upon 'mark of the mental,' we can and do classify various kinds of mentality: qualitative, intentional, and affective phenomena, for example, all fall under its scope. And we can beneficially address each of these individually. But if I am willing to address specific kinds of mental phenomena why am I not willing to address specific kinds of physical phenomena? Don't we also have a grasp of specific kinds of physical phenomena? As I've argued here and in more depth elsewhere, for the purpose of formulating the mind-body problem we do not (Montero 1999). With respect to the mental/nonmental distinction, while we do not have a definition of the mental, we nonetheless have a handle on the concept since we have a relatively clear idea of phenomena that fall on each side of the divide. However, with the physical/nonphysical distinction we lack even this. Of course, if panpsychism is true, everything will be fundamentally mental. But this does not mean that in stepping into the debate we have no grasp of the nonmental.[41] For we can easily conceive of something being fundamentally nonmental. Yet the concept of being fundamentally nonphysical seems to elude us entirely. What in the world (or, perhaps I should say 'out of the world') is supposed to count as being nonphysical?

The fact that we have no answer to this question shows that even if our grasp of the mental/nonmental distinction is far from clear, it is better than our grasp of the physical/nonphysical distinction. Furthermore, even if one thinks that both distinctions are equally opaque, this should not be reason to favour the current formulation of the mind-body problem because understanding the mental/nonmental distinction is no less exigent for understanding the current formulation than it is for understanding my proposed formulation. This is because those who think that the mind-body problem is the problem of explaining how the mind is physical assume that we have some intuitive understanding of that which they claim is entirely physical. (Beyond this they also assume an understanding of what it means to be physical.) More, of course, needs to be said about the mental. And, indeed, in debating the mind-body problem we are debating what exactly this should be.

VII: Spatiotemporality and Mentality

Some might object that I have missed my target entirely: the mind-body problem is not the question of whether mentality is fundamental, nor, for that matter, of whether it will ultimately be explainable by science (or, more specifically, physics). Rather, it is the question of whether it is spatiotemporal. For to be physical, some might say, is to be spatiotemporal.[42]

While this question echoes Descartes' concern with the mind-body problem—according to Descartes, mind is nonspatial, or at least unextended, and body is spatially extended—I think that it does not addresses the heart of the debate between physicalists and dualists.[43] For it seems to me that if mentality is fundamental, this, more than its being nonspatial or nonspatiotemporal, would capture what dualists believe is true about the mind: it is not reducible to anything else and thus has a rather special place in the world. What is more, being nonspatial or nonspatiotemporal seems neither sufficient nor necessary for dualism. For if the reason mentality is not spatiotemporal is simply that our theory of spacetime is incorrect I think that most physicalists would not take this to validate dualism. Or if mentality is in some sense purely abstract—the abstract program of the brain, perhaps—then physicalists who were happy with abstracta could be happy with a nonspatial mind.[44] Finally being nonspatiotemporal is not even necessary for dualism since if mentality were a *fundamental* spatiotemporal feature of the world, physicalists would not feel victorious. What matters to dualists is the fundamental nature of the mind, which is just what physicalists should argue against.

VIII: The Path Ahead

As I see it, then, focusing on such questions as whether mentality is a natural phenomenon, a physical phenomenon, or a spatial phenomenon sidesteps the hard question that lies at the heart of the debate. It is time to confront this question head-on: Is mentality a fundamental feature of the world? Physicalists will then need to make a decision: they can uphold deferential naturalism, the view, as Sellars put it, that science is the measure of all things, or they can put forth a substantive hypothesis about the general nature of the mind. To choose naturalism is to follow the course of science wherever it may lead,

which is, perhaps, not to abandon the mind-body problem, but to hand it over to someone else. But there is another option: take a stance and think of the mind-body problem as the problem of whether mentality is fundamentally nonmental. And without the cloak of naturalism, physicalists can do this openly and with a clear conscience. As such, they will not be naturalists in the sense that they will not simply be deferring to science to tell us what is and what is not. But, nevertheless, this does not mean that in putting forth their hypotheses they are necessarily being anti-scientific. To put forth a view, to state a hypothesis is to work hand in hand with science; to leave the mind-body problem up to someone else is not.

Of course, it may be the case that such hypotheses about the ultimate constituents of the universe might not admit of definitive refutation. For it might be difficult to know with certainty that any particular level is the bottom level. Alternatively, if there actually is no bottom level, if the world is in some sense infinitely divisible, then the question would become whether, after a certain level, it is nonmental *ad infinitum*. And who knows how to address that question. But in any case, I think that looking at the mind-body problem in terms of the distinction between the mental and the nonmental rather than the distinction between the physical and the nonphysical will not only relieve the conflict between naturalism and ontological significance (basically by giving each its own territory) but will also pave the way for what I hope will be a clearer, more interesting, and potentially even terminable debate about the fundamental nature of the mind.[45]

REFERENCES

Armstrong, D., "Naturalism, materialism, and first philosophy," in *Contemporary Materialism*, P. Moser and J. Trout, eds., (London: Routledge, 1995).

Brandom, R. *Making it Explicit: Reasoning, Representing and Discursive Commitment* (Cambridge, MA: Harvard University Press, 1994).

Chalmers, D. *The Conscious Mind* (Oxford: Oxford University Press, 1996).

Chomsky, N. *Language and Thought* (Wakefield, Rhode Island: Moyer Bell, 1993).

_____ "Language and nature," *Mind* 104 (1995): pp. 1–61.

_____ "Comments Galen Strawson, Mental Reality," *Philosophy and Phenomenological Research* LVIII (2) (June, 1968): pp. 437–41

Churchland, P. *Neurophilosophy: Toward a Unified Science of the Mind/Brain* (Cambridge, MA: MIT Press, 1995).

Clark, A. *Being There: Putting Brain, Body and World Together Again* (Cambridge, MA: MIT Press, 1997).

Crane, T., and H. Mellor, "There is no question of physicalism," *Mind* 99 (1990): pp. 185–206.

Daly, C., "What are physical properties?" *Pacific Philosophical Quarterly* 79 (1998): pp. 196–217.

Demopoulos, W., and Friedman, "Critical notice: Bertrand Russell's *The Analysis of Matter*: its historical and contemporary interest," *Philosophy of Science* 55 (1985): pp. 621–39.

Feigl, H., "The 'mental' and the 'physical,'" in *Concepts, Theories, and the Mind-Body Problem, Minnesota Studies in the Philosophy of Science*, vol. II, H. Feigl, M. Scriven, and G. Maxwell, eds., (Minneapolis: University of Minnesota Press 1958).

Fodor, J. *Psychosemantics* (Cambridge, MA: MIT Press, 1987).

Foster, J., "A defense of dualism," in *The Case for Dualism*, J. Smythies and J. Beloff, eds., (Charlottesville: University of Virginia Press, 1989).

Greene, B. *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory* (New York: W.W. Norton & Co., 1999).

Guttenplan, S., "An essay on mind," in *A Companion to Philosophy of Mind*, S. Guttenplan, ed., (Oxford: Basil Blackwell Ltd., 1995).

Healey, R., "Physicalists imperialism," *Proceedings of the Aristotelian Society* (1979): pp. 191–211.

Hellman, G., "Determination and logical truth," *Journal of Philosophy* 82 (1985): pp. 607–16.

Hornsby, J. *Simple Mindedness: In Defense of Naive Naturalism in the Philosophy of Mind* (Cambridge, MA: Harvard University Press, 1997).

Kim, J. *Philosophy of Mind* (Colorado: Westview Press, 1996).

_____ *Mind in a Physical World: An Essay on the Mind-body Problem and Mental Causation* (Cambridge, MA: MIT Press, 1998).

Kirk, R. *Raw Feeling: A Philosophical Account of the Essence of Consciousness* (New York: Oxford University Press, 1994).

Levine, J., "Conceivability and the metaphysics of mind," *Nôus* 32 (1998): pp. 449–80.

Lewis, D., "New work for a theory of universals," *Australasian Journal of Philosophy* 61 (1983): pp. 343–77.

McGinn, C., "Can we solve the mind-body problem?' *Mind* 98 (1989): pp. 349–66.

_____ "Consciousness and space," *Journal of Consciousness Studies* 2 (3) (1995): pp. 220–30.

Meehl, P., and W. Sellars, "The concept of emergence," in *The Foundations of Science and the Concept of Psychology and Psychoanalysis, Minnesota Studies in the Philosophy of Science*, Vol. I, H. Feigl and M. Scriven, ed., (Minneapolis: University of Minnesota Press, 1956).

Melnyk, A., "How to keep the 'physical' in physicalism," *Journal of Philosophy* 94 (1997): pp. 622–37.

Montero, B., "The body problem," *Nôus* 33 (1999): pp. 183–200.

Nagel, T. *The Last Word* (Oxford: Oxford University Press, 1996).

O'Leary-Hawthorne, J., and J. McDonough, "Numbers, minds, and bodies: A fresh look at mind-body dualism," in *Philosophical Perspectives 12, Language, Mind, and Ontology*, J. Tomberlin, ed., (Atascandero, CA: Ridgeview, 1998).

Papineau, D. *Philosophical Naturalism* (Oxford: Blackwell, 1993).

Poland, J. *Physicalism: The Philosophical Foundations* (Oxford: Oxford University Press, 1994).

Post, J. F. *The Faces of Existence: An Essay in Nonreductive Metaphysics* (Ithaca, NY: Cornell University Press, 1987).

Putnam, H., "On properties," in his *Mathematics, Matter and Method: Philosophical Papers, vol. 1* (Cambridge: Cambridge University Press, 1970).

Quine, W. V. *Theories and Things* (Cambridge, MA: Harvard University Press, 1981).

Russell, B. *An Outline of Philosophy* (London: Routledge, 1927/1992).

Scheffler, I., "The new dualism: Psychological and physical terms," *Journal of Philosophy* 47 (1950): pp. 732–52.

Searle, J. *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992).

Sellars, W., "Empiricism and the philosophy of mind," in his *Science, Perception, and Reality* (London: Routledge & Kegan Paul Ltd., 1963).

Shear, J., ed., *Explaining Consciousness: The 'Hard Problem'* (Cambridge, MA: MIT Press, 1997).

Smart, J., "The content of physicalism," *Philosophical Quarterly* 28 (1978): pp. 339–41.

Snowdon, P., "On formulating materialism and dualism," in *Cause, Mind, and Reality*, J. Heil, ed. (Dordrecht: Kluwer Academic Press, 1989).

Sober, E., "Physicalism from a probabilistic point of view," *Philosophical Studies* 95 (1999): pp. 135–74.

Strawson, G. *Mental Reality* (Cambridge MA: MIT Press, 1994a).

_____ "The experiential and the non-experiential," in *The Mind-body Problem: A Guide to the Current Debate*, R. Warner and T. Szubka, eds., (Oxford: Blackwell, 1994b).

_____ *Real Materialism and Other Essays* (Oxford University Press, 2008).

Stroud, B., "The charm of naturalism," *Proceedings and Addresses of the APA* 70 (1996): pp. 43–55.

Stich, S. *Deconstructing the Mind* (Oxford: Oxford University Press, 1996).

Stubenberg, L. *Consciousness and Qualia* (Amsterdam: J. Benjamins, 1998).

van Fraassen, B., "Against naturalized epistemology," in *On Quine: New Essays*, P. Leonardi and M. Santambrogio, eds., (Cambridge: Cambridge University Press, 1995).

van Fraassen, B., "Science, materialism, and false consciousness," in *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, J. Kvanvig, ed., (Lanham, MD: Rowman and Littlefield, 1996).

Wimsatt, W., "Reductionism, levels of organization, and the mind-body problem," in *Consciousness and the Brain: Scientific and Philosophic Strategies*, G. Globus, G. Maxwell, and I. Savodnik, eds., (New York: Plenum, 1976).

_____ "The ontology of complex systems: Levels of organization, perspectives, and causal thickets." *Canadian Journal of Philosophy* supp. (1994): p. 20.

Yablo, S., "Concepts and Consciousness," *Philosophy and Phenomenological Research* 59 (1999): pp. 455–63.

NOTES

1. The term 'fundamental' can, if you like, stand for whatever dependence relation you prefer. That is, when I say that the mind-body problem is the question of whether mentality is fundamentally nonmental you can substitute the question of whether mentality is reducible to (or constituted by, or supervenient on, etc.) the nonmental. Of course, the various notions of dependence are not unproblematic themselves, and there is little agreement on what relations between the lower level physical phenomena and higher lever mental phenomena suffice for physicalism. But let us take one problem at a time: the problem I am concerned with here is here is not how to understand the dependence relation, but how to understand the dependence base.
2. While I use the term 'mentality' rather than the more specific term 'experience,' most of what I say is directed at those engaged in the debate about experience, since many of those writing about intentionality already focus on the intentional/non-intentional distinction rather than the physical/non-physical distinction. Fodor 1987 is a good example: 'if the semantic and intentional are real properties of things, it must be in virtue of their identity with (or maybe supervenience on?) properties that are themselves neither intentional nor semantic.' (Thanks to Joseph Levine for pointing this out to me.)
3. At the 1999 Robert S. Cohen Colloquium: *Naturalism and its Discontents*, Kim emphasized this as does Stroud 1996.
4. Russell 1927/1992, p. 78. In an interesting forthcoming paper Galen Strawson points out that Joseph Priestley made more or less the same point in 1777.

5. Searle 1992, p. 84. To be sure, Searle is also not satisfied with the current terminology used to describe the mind-body problem.
6. Sellars 1963, p. 173. Or as Quine 1981, puts it, 'it is within science itself, and not in some prior philosophy, that reality is to be identified and described' (p. 21).
7. Credit goes to William Wimsatt for the droll phrase 'nothing butists.'
8. See Greene 1999. Speculation about such nonspatial, nontemporal stuff (or perhaps it would be better to call it 'nonstuff') should also be a bit worrisome for those who define the abstract over the nonspatiotemporal—do we want to say that our spatial world emerges out of abstracta?
9. Even if the results of mathematics, if true, are necessarily true, an argument is only interesting if there is some step in it that is not immediately obvious to everyone. (Why bother publishing a proof that everyone already knows?) Perhaps certain sceptical hypotheses, such as the hypothesis that the world was created five minutes ago with all apparent evidence of an earlier creation in place, are also, even in principle, unfalsifiable. But while we could never have evidence that could show such a hypothesis to be mistaken, there would still be an objective difference between the two situations—God, as it were, could know that the hypothesis is false. But if being physical amounts to simply existing, it is not clear that physicalism would be falsifiable even for God. Interestingly enough, Quine 1981 seems to accept the triviality of physicalism. For as he says, 'if the physicist suspected there was any event that did not consist in a redistribution of the elementary states allowed for by his physical theory he would seek a way of supplementing his theory' (p. 98).
10. See Lewis 1983. Lewis tries to carve out a position that is not trivial by explaining the physical in terms of whatever a future physics, which is significantly similar to current physics, but much improved, will tell us about the world. While one would like some explanation of what counts as 'significantly similar' and 'much improved,' the main difficulty with this notion of the physical is that if some groundbreaking discovery is made and physics goes through a *major* revolution, resulting in it not being sufficiently similar to today's physics, physicalists would, most likely, not want to claim the new posits and laws of this physics as being nonphysical. For other attempts to solve the body problem see Hellman 1985; Papineau 1993; Poland 1994; Meehl and Sellars 1956; Melnyk 1997; Smart 1978; Snowdon 1989.
11. In Montero 1999, I present an in-depth argument for this point. Since the main focus of this paper is to present a new way of thinking about the mind-body problem in light of the view that we have no solution to the body problem, my discussion here of this point will be brief. However, in arguing here for my proposed focus on the mind-body problem, I will also be arguing against retaining our current focus on the question of whether mentality is physical.
12. I should mention that I am not the first to use the term, 'post-physicalism.' John Post suggests that 'a far happier name [for his non-reductive physicalism], surely, would be "post-physicalism".' See Post 1987, p. 18.
13. There are a few exceptions. For example, Strawson 1994a seems to believe that the there are aspects of the world that are purely mental yet physical. And moreover, he seems to think that all physicalists must hold this view (especially pp. 46–59). Also see Strawson 2008. And Searle 1992 seems to hold a similar view. While O'Leary-Hawthorne and McDonough 1998 explicitly say that 'if an ideal physics will have consciousness at the metaphysical ground-floor, property dualism is wrong' (p. 350).
14. This claim, according to Kim 1996, is implied by the principle of supervenience assuming that if there can be one purely mental being, there can be at least two that differ mentality. For the principle of supervenience states that if two beings are psychologically discernible, then they will be physically discernible and this, according to Kim, shows that (given the further assumption) they cannot, then, be purely mental. But the principle that there are no fundamental mental entities follows from the principle of supervenience plus the assumption that if there is one purely mental being there can be more than one purely mental being only given a further assumption: that the physical realm does not include fundamental mental entities, itself.
15. To be sure, if you see science as the enterprise that investigates the world in nonanthropocentric or nonmental terms, an enterprise that we think essentially began in the seventeenth century with Descartes, then the two views I am trying to delineate are not distinct. (Thanks to Thomas Nagel for drawing my attention to this.) If this is what physicalists mean by 'science,' making this explicit should have the same affect on the debate as focusing on the question of whether mentality is fundamental.
16. I should emphasize that I am not intending to defend Wigner's hypothesis but am merely using it as a rough example of how physics might incorporate fundamental mentality. Yet it is not at all clear how to interpret it; in particular, it is not at all clear what is meant by 'acts of pure consciousness.' Apart from the question of what 'consciousness' means in this context (does an imaging device count? does animal consciousness?), there is the question of what 'pure' means. Does it mean fundamental? If not, then Wigner's hypothesis could be true even though mentality would not be fundamental.
17. Again, this is not a defence of the view. One problem with this version of the anthropic principle, as Steven Weinberg has pointed out, is that while the existence of life may place some constraints on the energy of this state of the carbon nucleus, it does not constrain it entirely. Furthermore, even if it did, it is not clear that it would count as an *explanation.*
18. Perhaps physicalists will claim that no *true* physics will account for mentality in this way. Perhaps not, but then they are really just claiming that (it is true that) mentality is not fundamental.
19. If one defines science as whatever tells us about the purely structural and relational, and one also holds that relations are never fundamental, then one probably would hold that a scientific account of the mental suffices to show that it is not fundamental. Russell 1927/1992 seems to hold such a view of science. And the view of science in Chalmers 1996 is very similar. Russell later abandoned this view due to Max Newman's criticism

that if this is all physics tells us about the world the only nontrivial information about the world that physics provides is information about how many things there are. (See Demopoulos and Friedman 1985 for an excellent discussion of this topic.) Since Chalmers' picture of what science tells us about the world includes causation it may avoid Newman's objection.

20. Others have expressed similar views of the mind-body problem. Strawson 1994a,b and 2008 emphasizes the importance of focusing on the distinction between the mental and the nonmental, specifically on the experiential and the nonexperiential. However, he also relies on a notion of the physical. And Levine 1998 says that materialism in philosophy of mind is the thesis that 'there is no sharp discontinuity in nature between the mental and the nonmental' (p. 449).
21. Poland 1994 also discusses this problem and tries to resist it. While he explicitly rejects placing restrictions on physics (p. 159), he tries to give physicalism content by making a distinction between future physics (the course of which we cannot predict) and what physicists, in general, study: spacetime and the fundamental constituents of all occupants of space time and the fundamental attributes that account for all interaction of such occupants (pp. 163–4). Yet it seems that in putting forth a theory about what physicists study he is, nevertheless, restricting physics. I discuss the significance of spatiality for the mind-body problem in section VII.
22. Putnam 1970. Putnam addresses this problem by providing a programmatic definition of the fundamental magnitudes, that is, he defines them as those magnitudes which physicists currently take to be fundamental. While this may be of use for the purposes of his paper, I think that it does not work for the purposes of the mind-body problem since it leaves the physicalist asserting a view she thinks is more likely false than true. For further discussion of this point see Montero 1999.
23. While I am equating naturalism with deference to science, the term 'naturalism' is used in an enormous variety of ways. For example, Hornsby 1997 calls herself a 'naïve naturalist' even though she explicitly denies that the mind is amenable to scientific investigation, while Stich 1996 argues against 'naturalism' while defending a view he calls 'open ended pluralism' which seems to amount pretty much to deferential naturalism.
24. Much thanks to Gene Wittmere for his insightful comments on this issue. I should note that not all naturalists would be willing to go this route. For example, McGinn 1989 calls himself a naturalist yet he not only predicts, but also claims to have shown that mentality will never be accounted for in nonmental terms.
25. He explains these notions in Papineau 1993, pp. 10–16.
26. Papineau 1993, pp. 29–30. Of course, in order to avoid circularity, he also needs to explain what he means by '*physical* effect.' To do this, he relies on some 'paradigmatic physical effects,' of which he thinks we have an intuitive understanding. In Montero 1999, I argue that relying on intuitions in these sorts of cases will not work.
27. Some have argued that the naturalist's hypothesis itself is *a priori*. See van Fraassen 1995.
28. The official title of the Berkeley booklet is, the *Particle Physics Booklet*, an abridged version of the *Review of Particle Physics*. The information in these books can be found at http://pdg.lbl.gov/.
29. However, see Melnyk 1997 for an argument that we should ground physicalism in current physics.
30. As Nagel 1996 says, 'atheists have no more reason to be alarmed by fundamental and irreducible mind-world relations than by fundamental and irreducible laws of physics' (p. 131). David Chalmers 1996 has also argued for this point. As he sees it, the view that consciousness is a fundamental property is 'entirely compatible with a contemporary scientific worldview' (p. 127).
31. This is related to the claim one often hears that even if Mary in the black and white room knows all the dualistic facts, she still would not know what it is like to see red. The implication, then, is that if Jackson's thought experiment does pose a problem for physicalism, it poses just as serious a problem for dualism. But this is not quite right since dualists hold that the experience of seeing red is fundamental, something that cannot be explained in terms of anything else. The fact Mary would need to know, claims the dualist, is nothing less than what it is like to see red; and if she knows what it is like to see red, she knows what it is like to see red.
32. As I see it, Chalmers' 1996 arguments for dualism mainly fall out of his assumption that the intrinsic/extrinsic divide is sharp and that the physical is extrinsic while the mental is intrinsic. Indeed, given these assumptions, it seems that his arguments for the possibility of zombies are inessential. (Cf. Yablo 1999 claims that 'almost everything' in Chalmers' argument turns on the claim that zombie worlds are possible.)
33. See Brandom 1994, who argues that mentality presupposes norms and that norms, while they do not presuppose mentality, presuppose sociality.
34. I should again emphasize that I do not take the distinction between the mental and the nonmental to be exhaustive. The most interesting positions with respect to the normativity of the mental might be those that take mentality to be normative and normativity to be not quite mental but not quite nonmental either (depending on what we say about the social, Brandom may fall into this camp). As I said before, I do not think that the existence of such borderline cases reduces the usefulness of setting out the two extremes.
35. Kim 1996, p. 147. I should also point out that causal closure can mean (1) any physical effect (that has a cause) has a sufficient physical cause or (2) any physical cause has only physical effects, or both. But Kim's quote only addresses (1).
36. Further problems arise if being fundamentally mental implies being free, since freedom of the will may not be compatible with any type of overarching theory. However, it is not clear that being fundamentally mental does imply being free. Furthermore, the debate between compatibilists and incompatibilists is far from settled: free will might fit into the otherwise nonfree world in an intelligible way. Or perhaps, if Nancy Cartwright is correct, what we would normally think of as the nonmental world, is, in a sense, actually more free than not; on her view, most of the world is not law governed (neither deterministically nor probabilistically.) If so, fundamental

(free) mentality would be no more of a threat to a complete physics than most nonmental phenomena.

37. I thank Leopold Stubenberg for his comments on this point.
38. Of course, if C-fibre stimulation is thought of as fundamentally mental, as, interestingly enough, Feigl 1958 seems to think, eliminativism does not follow. See Stubenberg 1998 for an illuminating discussion and defence of Feigl's identity theory. As Stubenberg puts it, on Feigl's view, 'the brain is made of qualia.'
39. As Gabby Sakamoto said to me, one might think that just as there is a body problem from those who ask 'is mentality fundamentally physical?' there is a mind problem for those who ask 'is mentality fundamentally nonmental?' Sober 1999 makes this point as well.
40. Strawson 1994a remarks, only half-jokingly, that perhaps the best explanation for those who resolutely deny qualitative experience is that there really are zombies (functional duplicates of human beings that have no phenomenal states) and that the eliminativists are among them.
41. Some arguments for idealism, however, do intend to show this.
42. For example, according to Meehl and Sellars 1956: 'an event or entity is *physical*$_1$ if it belongs in the space-time network.' (Something is physical$_2$, they say, 'if it is definable in terms of theoretical primitives adequate to describe completely the actual states though not necessarily the potentialities of the universe before the appearance of life.' As I see it, the problem here is that if physicists discovered that some sort of life-force was created in the big bang, or, perhaps, that the big-bang theory was wrong and that sentient life has existed all along, then any minimal class of theoretical primitives adequate to describe the universe before the appearance of life is, vacuously, empty. And so the concept of physical$_2$ in this case would pick out those things definable from nothing.) See also Armstrong 1995, who defines naturalism as 'the doctrine that reality consists of nothing but a single all-embracing spatio-temporal system' (p. 35).
43. This is not to say that the question of whether mentality is spatial is not an interesting one (see, for example, McGinn 1995).
44. Of course, it is not easy to formulate the abstract/concrete distinction either.
45. I would like to thank Anne Eaton, Michael Forster, Joseph Goguen, Joel David Hamkins, John Haugeland, Sean Kelley, Thomas Nagel, Marya Schectman, Bradford Skow, Leopold Stubenberg, Michael Thompson, Michael Voytinsky, William Wimsatt, Gene Wittmere, and the three anonymous *JCS* referees for their very helpful comments.

2 Consciousness

Consciousness poses arguably the most difficult problems in the philosophy and science of the mind. What is consciousness? Can consciousness be explained in physical terms? Is consciousness itself something physical? How can we have a theory of consciousness? These questions have been central in the history of the philosophy of mind. Many of the foundational papers in Part I were concerned with questions about consciousness; but in recent years, these questions have received particularly close attention. Some of this discussion is represented by the papers in this section.

A. General

The term "consciousness" can be used with many different meanings, so at the start it is helpful to get these meanings clear. Ned Block's contribution to this section (chapter 23) is particular useful here. Block distinguishes a number of senses of "consciousness." The central sort of consciousness is *phenomenal consciousness*. We can say that a subject is phenomenally conscious when there is *something it is like* to be that subject, and a mental state is phenomenally conscious when there is something it is like to be in that state. Phenomenally conscious mental states include the experience of seeing colors, feeling pains, and experiencing mental images and emotions. All of these involve a certain qualitative, experiential character. Block distinguishes phenomenal consciousness from *access consciousness* (which involves a certain sort of access to information), *self-consciousness* (which involves representation of oneself), and *monitoring consciousness* (which involves representation of one's own mental states). All of these notions of consciousness are important, but phenomenal consciousness is by far the most puzzling, and it is the focus of most of the articles in this section.

Block's characterization of phenomenal consciousness in terms of "what it is like" is taken from Thomas Nagel's important 1974 paper, "What it is like to be a bat?" (chapter 24). Here, Nagel focuses attention on this aspect of consciousness and argues that it is particularly difficult to explain. A major source of the difficulty is that standard explanations in science and philosophy are cast in objective terms, but consciousness is subjective by its nature. We might know all about the objective functioning in a bat's brain, but we still would not know what it is like to be the bat, from its own subjective viewpoint. Nagel does not rule out an eventual understanding of consciousness in physical terms, but suggests that it may require conceptual developments that are as yet quite beyond us.

A quite different perspective is given by Daniel Dennett's paper "Quining Qualia" (chapter 25). The term "qualia," as standardly used, refers to the properties of mental states that characterize what it is like to have them; so qualia and phenomenal consciousness are tightly bound together. Dennett argues that there is no reason to believe that qualia exist. He suggests that qualia are standardly taken to be ineffable, intrinsic, private, and directly apprehensible; and he argues through a series of thought-experiments that there is no reason to believe that mental states have properties of this sort. Dennett suggests that the notion of "qualia" reflects a confusion and refers to no properties at all.

What sorts of *theories* of consciousness can we develop? Some such theories of consciousness are scientific theories, cast in terms of neural and/or computational mechanisms. Other such theories are philosophical theories, explaining consciousness in terms of other concepts that do not presuppose consciousness. The remaining papers in this section discuss three important philosophical theories: *higher-order thought* theories, *representationalist* theories, and *illusionist* theories.

Higher-order theories of consciousness explain conscious states in terms of the existence of a higher-order mental state that is directed at the original state. These theories can be divided into *higher-order perception* theories and *higher-order thought* theories, of which the latter are the more popular. David Rosenthal (chapter 26) has developed a detailed account of this sort, holding that a mental state is conscious when it is the object of an appropriate sort of higher-order thought. Rosenthal argues for this conclusion from the premise that conscious states are states that we are conscious of, develops the view in some depth, and draws out consequences.

Representationalist theories hold that consciousness is reducible to a sort of representation: to a first approximation, the nature of a conscious state is exhausted by the way in which it represents the world. Michael Tye (chapter 27) argues for representationalism, holding that perceptual consciousness is exhausted by its representational content. Tye argues for this conclusion in part by considering the "transparent" world-directed nature of conscious experience. Tye rebuts various objections and draws out a number of consequences of his view, including the consequence that the character of experience is not fully determined by the internal state of the subject.

Illusionist theories of consciousness hold that phenomenal consciousness is an illusion. In Keith Frankish's analysis (chapter 28), weak illusionism holds that some apparent features of phenomenal consciousness are illusory (but consciousness still exists), while strong illusionism holds that the very existence of phenomenal consciousness is illusory. Strong illusionism is highly counterintuitive, but Frankish suggests that it may be the best theory of consciousness that we can find.

FURTHER READING

Block, Flanagan, and Güzeldere 1997 is an excellent collection of important philosophical articles on consciousness. Baars and Newman 2001 does the same for important scientific articles on consciousness. Nagel 1986, Dennett 1991, Tye 1995, and Rosenthal 2005 give book-length statements of their respective views, while Frankish 2017 is a symposium on illusionism in response to his article. Numerous responses to Block's article are contained in the 1995 issue of *Behavioral and Brain Sciences* in which Block's article was published; some of these are included in the Block, Flanagan, and Güzeldere volume. Akins 1993 gives an account of the phenomenology of bats, based on empirical studies. Seager 1999 contains an extensive discussion of Dennett's views, while also giving a general introduction to philosophical issues about consciousness.

Another important higher-order thought view, on which consciousness involves *potential* higher-order thoughts, is set out by Carruthers 2000. Armstrong 1968 and Lycan

1995 develop higher-order perception views. Block (chapter 23), Byrne 1997, Dretske 1993, Güzeldere 1997, and Siewert 1998 present objections to higher-order views. Other representationalist views are laid out by Byrne 2001, Dretske 1995, Harman 1989, Lycan 1996, and Siewert 1998. Criticisms are developed by Block 1990, Neander 1998, and Warfield 1999.

Akins, K., "What is it like to be boring and myopic?" in *Dennett and his Critics*, B. Dahlbom, ed., (Cambridge, MA: Blackwell, 1993).

Baars, B. J., and Newman, J. *Essential Sources in the Scientific Study of Consciousness* (Cambridge, MA: MIT Press, 2001).

Block, N., Flanagan, O., and Güzeldere, G., eds., *The Nature of Consciousness: Philosophical Debates* (Cambridge, MA: MIT Press, 1997).

Block, N., "Inverted earth," *Philosophical Perspectives* 4 (1990): pp. 53–79.

Byrne, A., "Some like it HOT: consciousness and higher-order thoughts," *Philosophical Studies* 2 (1997): pp. 103–29.

_____ "Intentionalism defended," *Philosophical Review* 110:2 (2001).

Carruthers, P. *Phenomenal Consciousness: A Naturalistic Theory* (Cambridge: Cambridge University Press, 2000).

Chalmers, D. J. *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996).

Dennett, D.C. *Consciousness Explained* (Boston: Little-Brown, 1991).

Dretske, F., "Conscious experience," *Mind* 102 (1993): pp. 263–83.

_____ *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995).

Frankish, K. *Illusionism as a Theory of Consciousness* (Exeter, UK: Imprint Academic, 2017).

Güzeldere, G. "Is consciousness the perception of what passes in one's own mind?" in *Conscious Experience*, T. Metzinger, ed., (Exeter, UK: Imprint Academic, 1995).

Harman, G., "The intrinsic quality of experience," *Philosophical Perspectives* 4 (1990): pp. 31–52. Reprinted in N. Block, O. Flanagan, and G. Güzeldere, eds., *The Nature of Consciousness: Philosophical Debates*.

Lycan, W. G., "Consciousness as internal monitoring, I," *Philosophical Perspectives* 9 (1995): pp. 1–14. Reprinted in N. Block, O. Flanagan, and G. Güzeldere, eds., *The Nature of Consciousness: Philosophical Debates*.

Lycan, W. G. *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996).

Nagel, T. *The View from Nowhere* (New York: Oxford University Press 1986).

Neander, K., "The division of phenomenal labor: A problem for representationalist theories of consciousness," *Philosophical Perspectives* 12 (1998): pp. 411–34.

Peacocke, C. *Sense and Content: Experience, Thought, and their Relations* (New York: Oxford University Press, 1983).

Rosenthal, D. *Consciousness and Mind* (Oxford: Oxford University Press, 2005).

Seager, W. E. *Theories of Consciousness: An Introduction and Assessment* (London: Routledge, 1999).

Siewert, C. *The Significance of Consciousness* (Princeton, NJ: Princeton University Press, 1998).

Tye, M. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995).

_____ *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000).

Warfield, T., "Against representational theories of consciousness," *Journal of Consciousness Studies* 6 (1999): pp. 66–69.

B. Consciousness and Materialism

The most widely debated question in the philosophy of consciousness is whether consciousness is physical or nonphysical. The majority of contemporary philosophers of mind are materialists, but there have been a number of important arguments against materialism.

My paper (chapter 29) gives an overview of issues about the metaphysics of consciousness. (As such, it can be used to complement this introduction.) It distinguishes the "hard" and "easy" problems of consciousness, and summarizes the three main sorts of arguments against materialism about consciousness: conceivability arguments, knowledge arguments, and explanatory arguments. It divides the theoretical landscape of views according to how they react to these arguments, all of which proceed by establishing an *epistemic gap* between physical processes and consciousness and inferring an *ontological gap*. The paper distinguishes three sorts of broadly materialist views: type-A materialism (which denies the epistemic gap), type-B materialism (which accepts the epistemic gap but denies the ontological gap), and type-C materialism (which accepts the epistemic gap but holds that it can be closed in principle). It also distinguishes three sorts of broadly nonreductive views: type-D dualism (interactionism), type-E dualism (epiphenomenalism), and type-F monism (a sort of "panprotopsychism" locating the grounds of experience in the unknown intrinsic qualities of the physical world). In addition to taxonomy, the paper argues against the first three views and defends the last three views; as such, it can be seen as an extended argument against materialism about consciousness and an exploration of the alternative options.

Perhaps the most important arguments against materialism have been the *knowledge argument* and the *modal argument*. The knowledge argument argues that there are truths about consciousness that cannot be deduced from physical truths and infers that consciousness is nonphysical. Its classic statement is given by Frank Jackson's paper in this section (chapter 30). Jackson imagines a future neuroscientist, Mary, who has been brought up in a black-and-white room, but who knows all the physical truths about the brain. Jackson argues that although Mary knows all the physical facts, she does not know all the facts about consciousness: in particular, she does not know what it is like to see red. This can be seen by noting that she gains knowledge of what it is like to see red when she leaves the room. He infers that facts about consciousness are nonphysical facts, and that materialism is false and epiphenomenalism about consciousness is true. This paper is supplemented by an excerpt from Jackson's 1987 paper "What Mary Didn't Know," which gives a formalization of the knowledge argument.

Materialists have responded to this argument in a number of different ways. Almost everyone agrees that Mary learns *something* when she leaves the room, but materialists argue that this new knowledge does not threaten materialism. One important strategy (taken by type-A materialists in the sense above) argues that Mary does not gain any new *factual* knowledge, but merely gains an ability, analogous to the ability to ride a bicycle (which arguably goes beyond knowledge of facts). David Lewis (chapter 31) takes this strategy (the so-called "ability analysis"), holding that Mary lacked no factual knowledge when inside her black-and-white room, so that her physical knowledge was complete.

Another important group of arguments exploit the *modal* notions of possibility and necessity to argument against materialism. The classic argument of this sort is Descartes' argument in the Sixth Meditation (chapter 1) that he can imagine himself existing without a body, so it is possible that he could exist without a body, so he is not physical. Contemporary modal arguments do not embrace exactly this reasoning, but like Descartes' argument, they often proceed from the *conceivability* (or imaginability, or apparent possibility) of a dissociation between consciousness and physical processes to the possibility of such a dissociation, and so to the falsity of materialism. One argument of this sort, Chalmers (chapter 29), appeals to the conceivability of *zombies*: creatures that are physically identical to conscious beings, but that are not conscious. From here, the argument infers the possibility of zombies, and so the falsity of materialism.

At the end of his book *Naming and Necessity*, Saul Kripke gives a somewhat more intricate modal argument. Earlier in the book, Kripke has argued that true identities (such as "heat is the motion of molecules") are necessary and hold in all possible worlds. He argues that when it seems that we can imagine that that heat is the motion of molecules, what we are

really imagining is that something other than molecular motion could have the *appearance* of heat. In the selection reproduced here (chapter 32), Kripke argues that for any mental state (e.g. pain) and any physical state (e.g. C-fiber firing), we can imagine the mental state without the physical state and vice versa. Further, he argues that this cannot be explained away as merely involving the appearance of the mental state, since the appearance of pain is itself pain. So he argues that it is possible that the mental state can exist without the physical state and vice versa. If so, and if true identities are necessary, mental states cannot be identical to physical states. Kripke uses this as an argument against (type- and token-) identity theories, but related arguments can be made against materialism more generally.

Katalin Balog (chapter 33) gives a materialist response to modal arguments, focusing on the special role of *phenomenal concepts* in grounding our acquaintance with consciousness. This is a version of the *phenomenal concept strategy*, which aims to account for the epistemic gap between physical processes and consciousness in psychological terms. If this strategy succeeds, the materialist can acknowledge a gap between physical and phenomenal *concepts*, while denying any gap between physical and phenomenal *properties* in the world.

Hedda Hassel Mørch (chapter 34) offers a defense of *panpsychism*: the view that there is an element of consciousness in all matter. She argues that the hard problem of consciousness is paralleled by a *hard problem of matter:* we only know physical processes by their structure, and we do not know their intrinsic nature. The panpsychist suggests that the intrinsic nature of matter involves consciousness itself. This view accommodates the epistemological gap between physical processes and consciousness without requiring a dualism where consciousness is separate from the physical world.

FURTHER READING

Ludlow, Stoljar, and Nagasawa 2004, Alter and Walter 2007, and Coleman 2019 are two collections of articles on the knowledge argument and related topics. The ability hypothesis was first put forward by Nemirow 1990 and is argued against by Lycan 1995. Versions of the type-B response are put forward by Churchland 1985, Horgan 1984, Loar 1990, Lycan 1995, and others. Jackson (chapter 19) can be seen as a rejoinder to this sort of response. Jackson 1998 rejects the conclusion of the argument and embraces a sort of type-A view. Nordby 1990 is a real-life exemplar of the Mary situation.

Levine 1983 addresses Kripke's modal argument and argues for an explanatory gap involving consciousness. Loar 1990 and Hill 1997 address modal arguments in a manner broadly compatible with Balog's approach. In the literature, modal arguments involving zombies are put forward by Campbell 1970 and Kirk 1974, and are developed by Chalmers 1996. Gendler and Hawthorne 2002 contains a number of recent papers discussing modal arguments. Mørch's view about matter is grounded in some ideas of Russell 1926, which are developed in more depth by Lockwood 1989 and criticized by Foster 1991. Bruntrup and Jaskolla 2017 and Strawson 2007 are collections of articles on panpsychism, while Rosenberg 2004 and Goff 2019 are accessible books on the topic. McGinn 1989 argues that the problem of consciousness is unsolvable.

Alter, T., and Walter, S. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism* (New York: Oxford University Press, 2007).

Bruntrup, G., and Jaskolla, L. *Panpsychism: Contemporary Perspectives* (New York: Oxford University Press, 2017).

Campbell, K. K. *Body and Mind* (New York: Doubleday, 1970).

Chalmers, D. J. *The Conscious Mind* (New York: Oxford University Press, 1996).

_____ "Materialism and the metaphysics of modality," *Philosophy and Phenomenological Research* 59 (1999): pp. 473–93. [consc.net/papers/modality.html]

Churchland, P. M., "Reduction, qualia and the direct introspection of brain states," *Journal of Philosophy* 82 (1985): pp. 8–28. Reprinted in *A Neurocomputational Perspective* (Cambridge, MA: MIT Press, 1989).

Foster, J. *The Immaterial Self: A Defense of the Cartesian Dualist Conception of Mind* (London: Routledge, 1991).

Gendler, T., and Hawthorne, J. *Imaginability, Conceivability, and Possibility* (Oxford: Oxford University Press, 2002).

Goff, P. *Galileo's Error: Foundations for a New Science of Consciousness* (New York: Pantheon, 2019).

Hill, C., "Imaginability, conceivability, and the mind-body problem," *Philosophical Studies* 87 (1997): pp. 61–85.

Horgan, T., "Jackson on physical information and qualia," *Philosophical Quarterly* 34 (1984): pp. 147–83.

Jackson, F., "Postscript to 'What Mary didn't know,'" in *Mind, Method, and Conditionals* (Cambridge: Cambridge University Press, 1998).

Kirk, R., "Zombies vs materialists," *Proceedings of the Aristotelian Society*, Supplementary Volume 48 (1974): pp. 135–52.

Levine, J., "Materialism and qualia: The explanatory gap," *Pacific Philosophical Quarterly* 64 (1983): pp. 354–61.

Loar, B., "Phenomenal states," *Philosophical Perspectives* 4 (1990): pp. 81–108.

Lockwood, M. *Mind, Brain, and the Quantum* (Oxford: Oxford University Press, 1989).

Ludlow, P., Stoljar, D., and Nagasawa, Y., eds., *There's Something About Mary* (Cambridge, MA: MIT Press, 2004).

Lycan, W. G., "A limited defense of phenomenal information," in *Conscious Experience*, T. Metzinger, ed., (Exeter: Imprint Academic, 1995).

McGinn, C., "Can we solve the mind-body problem?" *Mind* 98 (1989): pp. 349–66.

Nemirow, L., "Physicalism and the cognitive role of acquaintance," in *Mind and Cognition*, W. Lycan, ed., (Oxford: Blackwell, 1990).

Nordby, K., "Vision in a complete achromat: A personal account," in *Night Vision: Basic, Clinical, and Applied Aspects*, R. Hess, L. Sharpe, and K. Nordby, eds., (Cambridge, UK: Cambridge University Press, 1990).

Rosenberg, G. H. *A Place for Consciousness: Probing the Structure of the Natural World* (New York: Oxford University Press, 2004).

Russell, B. *The Analysis of Matter* (London: Kegan Paul, 1927).

Strawson, G., ed., *Consciousness and its Place in Nature: Does Physicalism Entail Panpsychism?* (Exeter: Imprint Academic, 2007).

Concepts of Consciousness[1]

Ned Block

The concept of consciousness is a hybrid or better, a mongrel concept: the word 'consciousness' connotes a number of different concepts and denotes a number of different phenomena. We reason about 'consciousness using some premises that apply to one of the phenomena that fall under 'consciousness,' other premises that apply to other 'consciousness' and we end up with trouble. There are many parallels in the history of science. Aristotle used 'velocity' sometimes to mean average velocity and sometimes to mean instantaneous velocity; his failure to see the distinction caused confusion. The Florentine Experimenters of the 17th Century used a single word (roughly translatable as 'degree of heat') for temperature and for heat, generating paradoxes. For example, when they measured 'degree of heat' by whether various heat sources could melt paraffin, heat source A came out hotter than B, but when they measured 'degree of heat' by how much ice a heat source could melt in a given time, B was hotter than A.[2] These are very different cases, but there is a similarity, one that they share with the case of 'consciousness.' The similarity is: very different concepts are treated as a single concept. I think we all have some tendency to make this mistake in the case of 'consciousness.'

Phenomenal Consciousness

First, consider phenomenal consciousness, or P-consciousness, as I will call it. Phenomenal consciousness is experience; what makes a state phenomenally conscious is that there is something 'it is like' (Nage 1974) to be in that state. Let me acknowledge at the outset that I cannot define P-consciousness in any remotely non-circular way. I don't consider this an embarrassment. The history of reductive definitions in philosophy should lead one not to expect a reductive definition of anything. But the best one can do for P-consciousness is in some respects worse than for many other things because really all one can do is *point* to the phenomenon (cf. Goldman 1993a). Nonetheless, it is important to point properly. John Searle, acknowledging that consciousness cannot be defined non-circularly, defines it as follows:

> By consciousness I simply mean those subjective states of awareness or sentience that begin when one wakes in the morning and continue throughout the period that one is awake until one falls into a dreamless sleep, into a coma, or dies or is otherwise, as they say, unconscious. [This comes from Searle 1990; there is a much longer attempt along the same lines in his 1992, p. 83ff.]

I will argue that this sort of pointing is flawed because it points to too many things, too many different consciousnesses.

So how should we point to P-consciousness? Well, one way is via rough synonyms. As I said, P-consciousness is experience. P-conscious properties are experiential properties. P-conscious states are experiential states; that is, a state is P-conscious just in case it has experiential properties. The totality of the experiential properties of a state are 'what it is like' to have it. Moving from synonyms to examples, we have P-conscious states when we see, hear, smell, taste, and have pains. P-conscious properties include the experiential properties of sensations, feelings, and perceptions, but I would also include thoughts, wants, and emotions.[3] An important feature of P-consciousness is that differences in intentional content often make a P-conscious difference. What it is like to hear a

Abridged and revised from "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences* 18 (1995): pp. 227–47, 1995, with the permission of Cambridge University Press.

sound as coming from the left differs from what it is like to hear a sound as coming from the right. Further, P-conscious differences often make an intentional difference. And this is partially explained by the fact that P-consciousness is often—perhaps even always—representational. (See Jackendoff 1987; van Gulick 1989; McGinn 1991, chapter 2; Flanagan 1992, chapter 4; Goldman 1993b.) So far, I don't take myself to have said anything terribly controversial. The controversial part is that I take P-conscious properties to be distinct from any cognitive, intentional, or functional property. At least, no such reduction of P-consciousness to the cognitive, intentional or functional can be known in the armchair manner of recent deflationist approaches. (Cognitive = essentially involving thought; intentional properties = properties in virtue of which a representation or state is about something; functional properties = e.g., properties definable in terms of a computer program. See Searle 1983 on intentionality; See Block 1980, 1994, for better characterizations of a functional property.) But I am trying hard to limit the controversiality of my assumptions. Though I will be assuming that functionalism about P-consciousness is false, I will be pointing out that limited versions of many of the points I will be making can be acceptable to the functionalist.[4]

By way of homing in on P-consciousness, it is useful to appeal to what may be a contingent property of it, namely the famous 'explanatory gap.' To quote T. H. Huxley (1866), 'How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of Djin when Aladdin rubbed his lamp.' Consider a famous neurophysiological theory of P-consciousness offered by Francis Crick and Christof Koch: namely, that a synchronized 35–75 hertz neural oscillation in the sensory areas of the cortex is at the heart of phenomenal consciousness. Assuming for the moment that such neural oscillations are the neural basis of sensory consciousness, no one has produced the concepts that would allow us to explain why such oscillations are the neural basis of one phenomenally conscious state rather than another or why the oscillations are the neural basis of a phenomenally conscious state rather than a phenomenally unconscious state.

However, Crick and Koch have offered a sketch of an account of how the 35–75 hertz oscillation might contribute to a solution to the 'binding problem.' Suppose one simultaneously sees a red square moving to the right and a blue circle moving to the left. Different areas of the visual cortex are differentially sensitive to color, shape, motion, etc. so what binds together redness, squareness and rightward motion? That is, why don't you see redness and blueness without seeing them as belonging with particular shapes and particular motions? And why aren't the colors normally seen as bound to the wrong shapes and motions? Representations of colors, shapes, and motions of a single object are supposed to involve oscillations that are in phase with one another but not with representations of other objects. But even if the oscillation hypothesis deals with the informational aspect of the binding problem (and there is some evidence against it), how does it explain *what it is like to see something as red in the first place*—or for that matter, as square or as moving to the right? Why couldn't there be brains functionally or physiologically just like ours, including oscillation patterns, whose owners' experience was different from ours or who had no experience at all? (Note that I don't say that there *could be* such brains. I just want to know *why not.*) No one has a clue how to answer these questions.

The explanatory gap in the case of P-consciousness contrasts with our better (though still not very good) understanding of the scientific basis of cognition. We have two serious research programs into the nature of cognition, the classical 'language of thought' paradigm, and the connectionist research program. Both assume that the scientific basis of cognition is computational. If this idea is right—and it seems increasingly promising—it gives us a better grip on why the neural basis of a thought state is the neural basis of that thought rather than some other thought or none at all than we have about the analogous issue for consciousness.

What I've been saying about P-consciousness is of course controversial in a variety of ways, both for some advocates and some opponents of some notion of P-consciousness. I have tried to steer clear of some controversies, e.g., controversies over inverted and absent qualia; over Jackson's (1986) Mary, the woman who is raised in a black and white room, learning all the physiological and functional facts about the brain and color vision, but nonetheless discovers a new fact when she goes outside the room for the first time and learns what it is like to see red; and even Nagel's view that we cannot know what it is like to be a bat.[5] Even if you think that P-consciousness as I have described it is an incoherent notion, you may be able to

agree with the main point of this paper, which is that a great deal of confusion arises as a result of confusing P-consciousness with something else. Not even the concept of what time it is now on the sun is so confused that it cannot itself be confused with something else.

Access-Consciousness

I now turn to the non-phenomenal notion of consciousness that is most easily and dangerously conflated with P-consciousness: access-consciousness. I will characterize access-consciousness, give some examples of how it makes sense for someone to have access-consciousness without phenomenal consciousness and vice versa, and then go on to the main theme of the paper, the damage done by conflating the two.

A-consciousness is access-consciousness. A representation is A-conscious if it is broadcast for free use in reasoning and for direct 'rational' control of action (including reporting). An A-state is one that consists in having an A-representation. I see A-consciousness as a cluster concept in which reportability is the element of the cluster that has the smallest weight even though it is often the best practical guide to A-consciousness.

The 'rational' is meant to rule out the kind of automatic control that obtains in blindsight. (Blindsight is a syndrome involving patients who have brain damage in the first stage of visual processing, primary visual cortex. These patients seem to have 'holes' in their visual fields. If the experimenter flashes stimuli in these holes and asks the patient what was flashed, the patient claims to see nothing but can often guess at high levels of accuracy, choosing between two locations or directions or whether what was flashed was an 'X' or an 'O'.)

I will suggest that A-consciousness plays a deep role in our ordinary 'consciousness' talk and thought. However, I must admit at the outset that this role allows for substantial indeterminacy in the concept itself. In addition, there are some loose ends in the characterization of the concept which cannot be tied up without deciding about certain controversial issues, to be mentioned below.[6] My guide in making precise the notion of A-consciousness is to formulate an information processing correlate of P-consciousness that is not ad hoc and mirrors P-consciousness as well as a non–ad hoc information processing notion can.

In the original version of this paper, I defined 'A-consciousness' as (roughly) 'poised for control of speech, reasoning, and action.'[7] In a comment on the original version of this paper, David Chalmers 1997 suggested defining 'A-consciousness' instead as 'directly available for global control.' Chalmers' definition has the advantage of avoiding enumerating the kinds of control. That makes the notion more general, applying to creatures who have kinds of control that differ from ours. But it has the disadvantage of that advantage, counting simple organisms as having A-consciousness if they have representations that are directly available for global control of whatever resources they happen to have. If the idea of A-consciousness is to be an information processing image of P-consciousness, it would not do to count a slug as having A-conscious states simply because there is some machinery of control of the resources that a slug happens to command.

As I noted, my goal in precisifying the ordinary notion of access as it is used in thinking about consciousness is to formulate a non–ad hoc notion that is close to an information processing image of P-consciousness. A flaw in both my definition and Chalmers' definition is that they make A-consciousness dispositional whereas P-consciousness is occurrent. As noted in the critique by Atkinson and Davies 1995, that makes the relation between P-consciousness and A-consciousness the relation between the ground of a disposition and the disposition itself. (See also Burge 1997.) This has long been one ground of criticism of both functionalism and behaviorism (Block and Fodor 1972), but there is no real need for an information-processing notion of consciousness to be saddled with a category mistake of this sort. I have dealt with the issue here by using the term 'broadcast,' as in Baars' 1988 theory that conscious representations are ones that are broadcast in a global workspace. A-consciousness is similar to that notion and to Dennett's 1993 notion of consciousness as cerebral celebrity.[8]

The interest in the A/P distinction arises from the battle between two different conceptions of the mind, the biological and the computational. The computational approach supposes that all of the mind (including consciousness) can be captured with notions of information processing, computation, and function in a system. According to this view (often called functionalism by philosophers), the level of abstraction for understanding the mind is one that allows multiple realizations, just as one computer can

be realized electrically or hydraulically. Their bet is that the different realizations don't matter to the mind, generally, and to consciousness specifically. The biological approach bets that the realization does matter. If P = A, the information processing side is right. But if the biological nature of experience is crucial, then realizations *do* matter, and we can expect that P and A will diverge.[9]

Although I make a distinction between A-consciousness and P-consciousness, I also want to insist that they interact. For example, what perceptual information is being accessed can change figure to ground and conversely, and a figure-ground switch can affect one's phenomenal state. For example, attending to the feel of the shirt on your neck, accessing those perceptual contents, switches what was in the background to the foreground, thereby changing one's phenomenal state. (See Hill 1991, 118–26; Searle 1992.)

Of course, there are notions of access in which the blindsight patient's guesses count as access. There is no right or wrong here. Access comes in various degrees and kinds, and my choice here is mainly determined by the desideratum of finding a notion of A-consciousness that mirrors P-consciousness. If the blindsight patient's perceptual representations are not P-conscious, it would not do to count them as A-conscious. (I also happen to think that the notion I characterize is more or less one that plays a big role in our thought, but that won't be a major factor here.)

I will mention three main differences between P-consciousness and A-consciousness. The first point, *put crudely*, is that P-conscious content is phenomenal, whereas A-conscious content is representational. It is of the essence of A-conscious content to play a role in reasoning, and only representational content can figure in reasoning. The reason this way of putting the point is crude is that many (perhaps even all) phenomenal contents are *also* representational. And some of the representational contents of a P-conscious state may be intrinsic to those P-contents.[10]

(In the last paragraph, I used the notion of P-conscious *content*. The P-conscious content of a state is the totality of the state's experiential properties, what it is like to be in that state. One can think of the P-conscious content of a state as the state's experiential 'value' by analogy to the representational content as the state's representational 'value.' In my view, the content of an experience can be both P-conscious and A-conscious; the former in virtue of its phenomenal feel and the latter in virtue of its representational properties.)

A closely related point: A-conscious states are necessarily transitive: A-conscious states must always be states of consciousness *of*. P-conscious states, by contrast, sometimes are and sometimes are not transitive. P-consciousness, as such, is not consciousness of. (I'll return to this point in a few paragraphs.)

Second, A-consciousness is a functional notion, and so A-conscious content is system-relative: what makes a state A-conscious is what a representation of its content does in a system. P-consciousness is not a functional notion.[11] In terms of Schacter's model of the mind (see the original version of this paper Block 1995), content gets to be P-conscious because of what happens *inside* the P-consciousness module. But what makes content A-conscious is not anything that could go on *inside* a module, but rather informational relations *among* modules. Content is A-conscious in virtue of (a representation with that content) reaching the Executive system, the system that is in charge of rational control of action and speech, and to that extent, we could regard the Executive module as the A-consciousness module. But to regard *anything* as an A-consciousness module is misleading, because what makes a typical A-conscious representation A-conscious is what getting to the Executive module sets it up to *do*, namely affect reasoning and action.

A third difference is that there is such a thing as a P-conscious *type* or *kind* of state. For example the feel of pain is a P-conscious type—every pain must have that feel. But any particular token thought that is A-conscious at a given time could fail to be accessible at some other time, just as my car is accessible now, but will not be later when my wife has it. A state whose content is informationally promiscuous now may not be so later.

The paradigm P-conscious states are sensations, whereas the paradigm A-conscious states are 'propositional attitude' states like thoughts, beliefs and desires, states with representational content expressed by 'that' clauses. (E.g., the thought that grass is green.) What, then, gets broadcast when a P-conscious state is also A-conscious? The most straightforward answer is: the P-content itself. However, exactly what this comes to depends on what exactly P-content is. If P-content is non-conceptual, it may be said that P contents are not the right sort of thing to play a role in inference and guiding action.

However, even with non-humans, pain plays a rational role in guiding action. Different actions are appropriate responses to pains in different locations. Since the contents of pain do in fact play a rational role, either their contents are conceptualized *enough*, or else nonconceptual or not very conceptual content can play a rational role.

There is a familiar distinction, alluded to above, between 'consciousness' in the sense in which we speak of a state as being a conscious state (intransitive consciousness) and consciousness *of* something (transitive consciousness). (The transitive/intransitive terminology seems to have appeared first in Malcolm 1984, but see also Rosenthal 1997. Humphrey 1992 mentions that the intransitive usage is much more recent, only 200 years old.) It is easy to fall into an identification of P-consciousness with intransitive consciousness and a corresponding identification of access-consciousness with transitive consciousness. Such an identification is over simple. As I mentioned earlier, P-conscious contents can be representational. Consider a perceptual state of seeing a square. This state has a P-conscious content that represents something, a square, and thus it is a state of P-consciousness *of* the square. It is a state of P-consciousness of the square even if it doesn't represent the square *as* a square, as would be the case if the perceptual state is a state of an animal that doesn't have the concept of a square. Since there can be P-consciousness *of* something, P-consciousness is not to be identified with intransitive consciousness.

Here is a second reason why the transitive/intransitive distinction cannot be identified with the P-consciousness/A-consciousness distinction: The *of*-ness required for transitivity does not guarantee that a content be utilizable by a **consuming** system, a system that uses the representations for reasoning or planning or control of action at the level required for A-consciousness. For example, a perceptual state of a brain-damaged creature might be a state of P-consciousness of, say, motion, even though connections to reasoning and rational control of action are damaged so that the state is not A-conscious. In sum, P-consciousness can be consciousness of, and consciousness of need not be A-consciousness.

Those who are uncomfortable with P-consciousness should pay close attention to A-consciousness because it is a good candidate for a reductionist identification with P-consciousness.[12]

Many of my critics (Searle 1992, Burge 1997) have noted that if there can be 'zombies,' cases of A without P, they are not conscious in any sense of the term. I am sympathetic, but I don't agree with the conclusion that some have drawn that the A-sense is not a sense of 'consciousness' and that A is not a kind of consciousness. A-consciousness can be a kind of consciousness even if it is parasitic on a core notion of P-consciousness. A parquet floor is a floor even though it requires another floor beneath it. A-consciousness can come and go against a background of P-consciousness.

The rationale for calling A-consciousness a kind of consciousness is first that it fits a certain kind of quasi-ordinary usage. Suppose one has a vivid mental image that is repressed. Repression need not make the image go away or make it non-phenomenal. One might realize after psychoanalysis that one had the image all along, but that one could not cope with it. It is 'unconscious' in the Freudian sense—which is A-unconsciousness. Second, A-consciousness is typically the kind of consciousness that is relevant to use of words like 'conscious' and 'aware' in cognitive neuroscience. This point is made in detail in my comment on a special issue of the journal *Cognition* (Block 2001) This issue summarizes the 'state of the art' and some of the writers are clearly talking about A-consciousness (or one or another version of monitoring consciousness—see below) whereas others are usually talking about P-consciousness. The A notion of consciousness is the most prominent one in the discussion in that issue and in much of the rest of cognitive neuroscience. (See the article by Dehaene and Naccache in that volume which is very explicit about the use of A-consciousness.) Finally, recall that my purpose in framing the notion of A-consciousness is to get a functional notion of consciousness that is not ad hoc and comes as close to matching P-consciousness as a purely functional notion can. I hope to show that nonetheless there are cracks between P and A. In this context, I prefer to be liberal with terminology, allowing that A is a form of consciousness but not identical to phenomenal consciousness.

A-Consciousness without P-Consciousness

The main point of this paper is that these two concepts of consciousness are distinct and quite likely have different extensions yet are easily

confused. Let us consider conceptually possible cases of one without the other. Actual cases will be more controversial.

First, I will give some putative examples of A-consciousness without P-consciousness. If there could be a full-fledged phenomenal zombie, say a robot computationally identical to a person, but whose silicon brain did not support P-consciousness, that would do the trick. I think such cases conceptually possible, but this is very controversial. (See Shoemaker 1975, 1981.)

But there is a less controversial kind of case, a very limited sort of partial zombie. Consider the blindsight patient who 'guesses' that there is an 'X' rather than an 'O' in his blind field. Taking his word for it (for the moment), I am assuming that he has no P-consciousness of the 'X'. The blindsight patient also has no 'X'-representing A-conscious content, because although the information that there is an 'X' affects his 'guess,' it is not available as a premise in reasoning (until he has the quite distinct state of hearing and believing his own guess), or for rational control of action or speech. Marcel 1986 points out that the thirsty blindsight patient would not reach for a glass of water in the blind field. So the blindsight patient's perceptual or quasi-perceptual state is unconscious in the phenomenal *and* access senses (*and* in the monitoring senses to be mentioned below too).

Now imagine something that may not exist, what we might call *superblindsight.* A real blindsight patient can only guess when given a choice from a small set of alternatives ('X'/'O'; horizontal/vertical, etc.). But suppose—interestingly, apparently contrary to fact—that a blindsight patient could be trained to prompt himself at will, guessing what is in the blind field without being told to guess. The superblindsighter spontaneously says, 'Now I know that there is a horizontal line in my blind field even though I don't actually see it.' Visual information of a certain limited sort (excluding color and complicated shapes) from his blind field simply pops into his thoughts in the way that solutions to problems we've been worrying about pop into our thoughts, or in the way some people just know the time or which way is North without having any perceptual experience of it. He knows there is an 'X' in his blind field, but he doesn't know the type font of the 'X.' The superblindsighter himself contrasts what it is like to know visually about an 'X' in his blind field and an 'X' in his sighted field. There is something it is like to experience the latter, but not the former he says. It is the difference between *just knowing* and knowing via a visual experience. Taking his word for it, here is the point: the perceptual content that there is an 'X' in his visual field is A-conscious but not P-conscious. The superblindsight case is a very limited partial zombie.

Of course, the superblindsighter has a *thought* that there is an 'X' in his blind field that is *both* A-conscious and P-conscious. But I am not talking about the thought. Rather, I am talking about the state of his perceptual system that gives rise to the thought. It is this state that is A-conscious without being P-conscious.[13]

The (apparent) non-existence of superblindsight is a striking fact, one that a number of writers have noticed, more or less. What Marcel was in effect pointing out was that the blindsight patients, in not reaching for a glass of water, are not superblindsighters. (See also Farah 1994.) Blind perception is never super blind perception.[14]

Notice that the superblindsighter I have described is just a little bit different (though in a crucial way) from the ordinary blindsight patient. In particular, I am *not relying* on what might be thought of as a full-fledged *quasizombie*, a super-*duper*-blindsighter whose blindsight is *every bit* as good, functionally speaking, as his sight. In the case of the superduper blindsighter, the *only* difference between vision in the blind and sighted fields, functionally speaking, is that the quasi-zombie himself regards them differently. Such an example will be regarded by some (though not me) as incoherent—see Dennett 1991, for example. But we can avoid disagreement about the super-duper-blindsighter by illustrating the idea of A-consciousness without P-consciousness by appealing only to the superblindsighter. Functionalists may want to know why the superblindsight case counts as A-conscious without P-consciousness. After all, they may say, if we have *really high-quality access* in mind, the superblindsighter that I have described does not have it, so he lacks *both* P-consciousness and really high-quality A-consciousness. The super-duper-blindsighter, on the other hand, *has* both, according to the functionalist, so in neither case, according to the objection, is there A-consciousness without P-consciousness.

One could put the point by distinguishing three types of access: (1) really high-quality access, (2) medium access and (3) poor access. The *actual* blindsight patient has poor access (he has to be prompted to guess), the superblindsight patient has medium access and the

super-duper blindsight patient—as well as most of us—has really high-quality access. The functionalist objector I am talking about identifies P-consciousness with A-consciousness of the really high-quality kind, whereas I am allowing A-consciousness with only medium access. (We agree in excluding low-quality access.) The issue, then, is whether the functionalist can get away with restricting access to high quality access. I think not. I believe that in some cases, normal phenomenal vision involves only medium access. The easiest case to see for yourself with is peripheral vision. If you wave a colored object near your ear, you will find that in the right location you can see the movement without having the kind of rich access that you have in foveal vision. For example, your ability to recover shape and color is poor.

Why isn't peripheral vision a case of A without P? In peripheral vision, we are both A and P conscious of the same features—e.g., motion but not color. But in superblindsight—so the story goes—there is no P-consciousness of the horizontal line. (He just knows.) I conclude that A without P is conceptually possible even if not actual.

P-Consciousness without A-Consciousness

Consider an animal that you are happy to think of as having P-consciousness for which brain damage has destroyed centers of reasoning and rational control of action, thus preventing A-consciousness. It certainly seems *conceptually possible* that the neural bases of P-consciousness systems and A-consciousness systems be distinct, and if they are distinct, then it is possible, at least conceptually possible, for one to be damaged while the other is working well. Evidence has been accumulating for twenty-five years that the primate visual system has distinct dorsal and ventral subsystems. Though there is much disagreement about the specializations of the two systems, it does appear that much of the information in the ventral system is much more closely connected to P-consciousness than information in the dorsal system (Goodale and Milner 1992). So it may actually be possible to damage A-consciousness without P-consciousness and perhaps even conversely.[15]

Further, one might suppose (Rey 1983, 1988; White 1987) that some of our own subsystems—say each of the two hemispheres of the brain—might themselves be separately P-conscious. Some of these subsystems might also be A-consciousness, but other subsystems might not have sufficient machinery for reasoning or reporting or rational control of action to allow their P-conscious states to be A-conscious; so if those states are not accessible to another system that does have adequate machinery, they will be P-conscious but not A-conscious.

Here is another reason to believe in P-consciousness without A-consciousness: Suppose that you are engaged in intense conversation when suddenly at noon you realize that right outside your window, there is—and has been for some time—a pneumatic drill digging up the street. You were aware of the noise all along, one might say, but only at noon are you *consciously aware* of it. That is, you were P-conscious of the noise all along, but at noon you are both P-conscious *and* A-conscious of it. Of course, there is a very similar string of events in which the crucial event at noon is a bit more intellectual. In this alternative scenario, at noon you realize not just that there is and has been a noise, but also that *you are now and have been hearing* the noise. In this alternative scenario, you get 'higher order thought' as well as A-consciousness at noon. So on the first scenario, the belief that is acquired at noon is that there is and has been a noise, and on the second scenario, the beliefs that are acquired at noon are the first one plus the belief that you are and have been hearing the noise. But it is the first scenario, not the second that interests me. It is a good case of P-consciousness without A-consciousness. Only at noon is the content of your representation of the drill *broadcast* for use in rational control of action and speech. (Note that A-consciousness requires being broadcast, not merely being available for use.)

In addition, this case involves a natural use of 'conscious' and 'aware' for A-consciousness and P-consciousness. 'Conscious' and 'aware' are more or less synonymous, so when we have one of them we might think of it as awareness, but when we have both it is natural to call that conscious awareness. This case of P-consciousness without A-consciousness exploits what William James 1890 called 'secondary consciousness' (at least I think it does; James scholars may know better), a category that he may have meant to include cases of P-consciousness without attention.

I have found that the argument of the last paragraph makes those who are distrustful of

introspection uncomfortable. I agree that introspection is not the last word, but it is the first word, when it comes to P-consciousness. The example shows the conceptual distinctness of P-consciousness from A-consciousness and it also puts the burden of proof on anyone who would argue that as a matter of empirical fact they come to the same thing.

A-consciousness and P-consciousness very often occur together. When one or the other is missing, we can often speak of unconscious states (when the context is right). Thus, in virtue of missing A-consciousness, we think of Freudian states as unconscious. And in virtue of missing P-consciousness, it is natural to describe the superblindsighter or the unfeeling robot or computer as unconscious. Lack of monitoring-consciousness in the presence of A and P is also sometimes described as unconsciousness. Thus Julian Jaynes describes Greeks as becoming conscious when—in between the time of the Iliad and the Odyssey, they become more reflective.

Flanagan 1992 criticizes my notion of A-consciousness, suggesting that we replace it with a more liberal notion of informational sensitivity that counts the blindsight patient as having access-consciousness of the stimuli in his blind field. The idea is that the blindsight patient has **some** access to the information about the stimuli in the blind field, and that amount of access is enough for access consciousness. Of course, as I keep saying, the notion of A-consciousness that I have framed is just one of a family of access notions. But there is more than a verbal issue here. The real question is what good is A-consciousness as I have framed it in relation to the blindsight issue? The answer is that in blindsight, the patient is supposed to **lack** 'consciousness' of the stimuli in the blind field. My point is that the blindsight lacks both P-consciousness and a kind of access (both medium- and high-level access in the terminology used earlier), and that these are easily confused. This point is not challenged by pointing out that the blindsight patient also has a lower level of access to this information.

The kind of access that I have built into A-consciousness plays a role in theory outside of this issue and in daily life. Consider the Freudian unconscious. Suppose I have a Freudian unconscious desire to kill my father and marry my mother. Nothing in Freudian theory requires that this desire be P-unconscious; for all Freudians should care, it might be P-conscious. What is the key to the desire being Freudianly unconscious is that it come out in slips, dreams, and the like, but *not* be freely available as a premise in reasoning (in virtue of having the unconscious desire) and that it not be freely available to guide action and reporting. Coming out in slips and dreams *makes it conscious in Flanagan's sense*, so that sense of access is no good for capturing the Freudian idea. But it is unconscious in my A-sense. If I can just tell you that I have a desire to kill my father and marry my mother (and not as a result of therapy), then it isn't an unconscious state in either Freud's sense or my A sense. Similar points can be made about a number of the syndromes that are often regarded as disorders of consciousness. For example, consider prosopagnosia, a syndrome in which someone who can see noses, eyes, etc., cannot recognize faces. Prosopagnosia is a disorder of A-consciousness, not P-consciousness and not Flanagan's informational sensitivity. We count someone as a prosopagnosic even when they are able to guess at better than a chance level who the face belongs to, so that excludes Flanagan's notion. Further, P-consciousness is irrelevant, and that excludes P-consciousness as a criterion. It isn't the presence or absence of a feeling of familiarity that defines prosopagnosia, but rather the patient not knowing who the person is whose face he is seeing or whether he knows that person.

I am finished sketching the contrast between P-consciousness and A-consciousness. In the remainder of this section, I will briefly discuss two cognitive notions of consciousness, so that they are firmly distinguished from both P-consciousness and A-consciousness.

Self-Consciousness

By this term, I mean the possession of the concept of the self and the ability to use this concept in thinking about oneself. A number of higher primates show signs of recognizing that they see themselves in mirrors. They display interest in correspondences between their own actions and the movements of their mirror images. By contrast, dogs treat their mirror images as strangers at first, slowly habituating. In one experimental paradigm, experimenters painted colored spots on the foreheads and ears of anesthetized primates, watching what happened. Chimps between ages 7 and 15 usually try to wipe the spot off (Povinelli 1994; Gallup 1982). Monkeys do not do this, according to published

reports as of 1994. (Since then, Hauser et al. 1995, have shown that monkeys can pass the test if the mark is salient enough.) Human babies don't show similar behavior until the last half of their second year. Perhaps this is a test for self-consciousness. (Or perhaps it is only a test for understanding mirrors; but what is involved in understanding mirrors if not that it is oneself one is seeing?) But even if monkeys and dogs have no self-consciousness, no one should deny that they have P-conscious pains, or that there is something it is like for them to see their reflections in the mirror. P-conscious states often seem to have a 'me-ishness' about them, the phenomenal content often represents the state as a state of me. But this fact does not at all suggest that we can reduce P-consciousness to self-consciousness, since such 'me-ishness' is the same in states whose P-conscious content is different. For example, the experience as of red is the same as the experience as of green in self-orientation, but the two states are different in phenomenal feel.[16]

Monitoring-Consciousness

The idea of consciousness as some sort of internal monitoring takes many forms. One notion is that of some sort of inner perception. This could be a form of P-consciousness, namely P-consciousness of one's own states or of the self. Another notion is often put in information-processing terms: internal scanning. And a third, metacognitive notion, is that of a conscious state as one that is accompanied by a thought to the effect that one is in that state.[17] Let us lump these together as one or another form of monitoring-consciousness. Given my liberal terminological policy, I have no objection to monitoring-consciousness as a notion of consciousness. Where I balk is at the idea that P-consciousness just is one or another form of monitoring-consciousness.

To identify P-consciousness with internal scanning is just to grease the slide to eliminativism about P-consciousness. Indeed, as Georges Rey 1983 has pointed out, ordinary laptop computers are capable of various types of self-scanning, but as he also points out, no one would think of their laptop computer as 'conscious' (using the term in the ordinary way, without making any of the distinctions I've introduced). Since, according to Rey, internal scanning is essential to consciousness, he concludes that the concept of consciousness is incoherent. If one regards the various elements of the mongrel concept that I have been delineating as elements of a single concept, then that concept is indeed incoherent and needs repair by making distinctions along the lines I have been suggesting. I doubt that the ordinary concept of consciousness is sufficiently determinate for it to be incoherent, though whether or not this is so is an empirical question about how people use words that it is not my job to decide. However that inquiry turns out, Rey's mistake is to trumpet the putative incoherence of the concept of consciousness as if it showed the incoherence of the concept of *phenomenal* consciousness.[18]

Rosenthal 1997 defines reflexive consciousness as follows: S is a reflexively conscious state of mine ↔ S is accompanied by a thought—arrived at non-inferentially and non-observationally—to the effect that I am in S. He offers this 'higher order thought' (HOT) theory as a theory of phenomenal consciousness. It is obvious that phenomenal consciousness without HOT and HOT without phenomenal consciousness are both *conceptually* possible. For examples, perhaps dogs and infants have phenomenally conscious pains without higher order thoughts about them. For the converse case, imagine that by bio-feedback and imaging techniques of the distant future, I learn to detect the state in myself of having the Freudian unconscious thought that it would be nice to kill my father and marry my mother. I could come to know—non-inferentially and non-observationally—that I have this Freudian thought even though the thought is not phenomenally conscious.

Rosenthal sometimes talks as if it is supposed to be a basic law of nature that phenomenal states and HOTs about them co-occur. That is a very adventurous claim. But even if it is true, then there must be a mechanism that explains the correlation, as the fact that both heat and electricity are carried by free electrons explains the correlation of electrical and thermal conductivity. But any mechanism breaks down under extreme conditions, as does the correlation of electrical and thermal conductivity at extremely high temperatures. So the correlation between phenomenality and HOT would break down too, showing that higher order thought does not yield the basic scientific nature of phenomenality.

Rosenthal's definition of his version of monitoring-consciousness has a number of ad hoc features. 'Non-observationally' is required to rule out (e.g.) a case in which I know about a thought I have repressed by observing my own

behavior. 'Non-inferentially' is needed to avoid a somewhat different case in which I appreciate (non-observationally) my own pain and infer a repressed thought from it. Further, Rosenthal's definition involves a stipulation that the possessor of the monitoring-conscious state is the same as the thinker of the thought—otherwise *my* thinking about *your* pain would make *it* a conscious pain. All these ad hoc features can be eliminated by moving to the following definition of monitoring-consciousness: S is a monitoring-conscious state ↔ S is phenomenally presented in a thought about S. This definition uses the notion of phenomenality, but this is no disadvantage unless one holds that there is no such thing apart from monitoring itself. The new definition, requiring phenomenality as it does, has the additional advantage of making it clear why monitoring-consciousness is a kind of *consciousness.*

There is an element of plausibility to the collapse of P-consciousness into monitoring-consciousness. Consider two dogs, one of which has a perceptual state whereas the other has a similar perceptual state plus a representation of it. Surely the latter dog has a conscious state even if the former dog does not. Quite right, because *consciousness of* plausibly brings consciousness with it. (I'm only endorsing the plausibility of this idea, not its truth.) But the *converse* is more problematic. If I am conscious of a pain or a thought, then, plausibly, that pain or thought has some P-conscious aspect. But even if consciousness of entails P-consciousness, that gives us no reason to believe that P-consciousness entails consciousness of, and it is the implausibility of this converse proposition that is pointed to by the dog problem. The first dog can have a P-conscious state too, even if it is not conscious of it.

Perhaps you are wondering why I am being so terminologically liberal, counting P-consciousness, A-consciousness, monitoring consciousness and self-consciousness all as types of consciousness. Oddly, I find that many critics wonder why I would count *phenomenal* consciousness as consciousness, whereas many others wonder why I would count access or monitoring or *self* consciousness as consciousness. In fact two reviewers of this paper complained about my terminological liberalism, but for incompatible reasons. One reviewer said: 'While what he uses ['P-consciousness'] to refer to—the 'what it is like' aspect of mentality—seems to me interesting and important, I suspect that the discussion of it under the heading 'consciousness' is a source of confusion . . . he is right to distinguish access-consciousness (which is what I think deserves the name 'consciousness') from this.' Another reviewer said: 'I really still can't see why access is called . . . access-consciousness? Why isn't access just . . . a purely information processing (functionalist) analysis?' This is not a merely verbal matter. In my view, all of us, despite our explicit verbal preferences, have some tendency to use 'conscious' and related words in both ways, and our failure to see this causes a good deal of difficulty in thinking about 'consciousness.'

I've been talking about different concepts of 'consciousness' and I've also said that *the* concept of consciousness is a mongrel concept. Perhaps, you are thinking, I should make up my mind. My view is that 'consciousness' is actually an ambiguous word, though the ambiguity I have in mind is not one that I've found in any dictionary. I started the paper with an analogy between 'consciousness' and 'velocity,' and I think there is an important similarity. One important difference, however, is that in the case of 'velocity,' it is easy to get rid of the temptation to conflate the two senses, even though for many purposes the distinction is not very useful. With 'consciousness,' there is a tendency towards 'now you see it, now you don't.' I think the main reason for this is that P-consciousness presents itself to us in a way that makes it hard to imagine how a conscious state could fail to be accessible and self-reflective, so it is easy to fall into habits of thought that do not distinguish these concepts.[19]

The chief alternative to the ambiguity hypothesis is that there is a single concept of consciousness that is a *cluster concept.* For example, a prototypical religion involves belief in supernatural beings, sacred and profane objects, rituals, a moral code, religious feelings, prayer, a worldview, an organization of life based on the world view and a social group bound together by the previous items (Alston 1967). But for all of these items, there are actual or possible religions that lack them. For example, some forms of Buddhism do not involve belief in a supreme being and Quakers have no sacred objects. It is convenient for us to use a concept of religion that binds together a number of disparate concepts whose referents are often found together.

The distinction between ambiguity and cluster concept can be drawn in a number of equally legitimate ways that classify some cases differently. That is, there is some indeterminacy in

the distinction. Some might even say that *velocity* is a cluster concept because for many purposes it is convenient to group average and instantaneous velocity together. I favor tying the distinction to the clear and present danger of conflation, especially in the form of equivocation in an argument. Of course, this is no analysis, since equivocation is definable in terms of ambiguity. My point, rather, is that one can make up one's mind about whether there is ambiguity by finding equivocation hard to deny. In Block 1995, the longer paper from which this paper derives, I give some examples of conflations.

When I called *consciousness* a mongrel concept I was not declaring allegiance to the cluster theory. Rather, what I had in mind was that an ambiguous word often corresponds to an ambiguous mental representation, one that functions in thought as a unitary entity and thereby misleads. These are mongrels. I would also describe *velocity* and *degree of heat* (as used by the Florentine Experiments of the 17th century) as mongrel concepts. This is the grain of truth in the cluster-concept theory.

Note the distinction between the claim that the concept of consciousness is a mongrel concept and the claim that consciousness is not a natural kind (Churchland 1983, 1986). The former is a claim about the concept, one that can be verified by reflection alone. The latter is like the claim that dirt or cancer are not natural kinds, claims that require empirical investigation.[20]

REFERENCES

Alston, W. "Religion," in *The Encyclopedia of Philosophy* (New York: Macmillan/Free Press, 1967), pp. 140–45.

Armstrong, D. M. *A Materialist Theory of Mind* (New York: Humanities Press, 1968).

_____ "What is consciousness?" in *The Nature of Mind* (Ithaca, NY: Cornell University Press, 1981).

Atkinson, A., and Davies, M., "Consciousness without conflation," *The Behavioral and Brain Sciences* 18, no. 2 (1995): pp. 248–49.

Baars, B. J. *A Cognitive Theory of Consciousness* (New York: Cambridge University Press, 1988).

Block, N., "What is functionalism?" in *Readings in the Philosophy of Psychology,* vol. 1, N. Block, ed., (Cambridge, MA: Harvard University Press, 1980).

_____ "Consciousness and accessibility," *Behavioral and Brain Sciences* 13 (1990): pp. 596–98.

_____ "Evidence against epiphenomenalism," *Behavioral and Brain Sciences* 14, no. 4 (1991): pp. 670–72.

_____ "Begging the question against phenomenal consciousness," *Behavioral and Brain Sciences* 15 (1992): pp. 205–6.

_____ "Review of D. *Dennett, Consciousness Explained, The Journal of Philosophy," XC*, no. 4 (1993): pp. 181–93.

_____ "'Consciousness,' 'Functionalism,' 'Qualia,'" in *A Companion to Philosophy of Mind,* S. Guttenplan, ed., (Cambridge, UK: Blackwell, 1994).

_____ "On a Confusion about a function of consciousness," *The Behavioral and Brain Sciences* 18 (1995): pp. 227–47.

_____ "Paradox and cross purposes in recent work on consciousness," *Cognition* 79 (2001): pp. 1–2, 197–219.

Block, N., and Dworkin, G., "IQ, heritability and inequality, Part I," *Philosophy and Public Affairs* 3, no. 4 (1974): pp. 331–409.

Burge, Tyler, "Two kinds of consciousness," in *The Nature of Consciousness: Philosophical Debates,* N. Block et al., eds., (Cambridge, MA: MIT Press, 1997).

Carruthers, P., "Brute experience," *Journal of Philosophy* 86 (1989): pp. 258–69.

_____ "Consciousness and concepts," *Proceedings of the Aristotelian Society,* Supplementary Volume LXVI (1992): pp. 40–59.

Chalmers, D. J., "Availability: The cognitive basis of experience?" in *The Nature of Consciousness,* N. Block et al., eds. (Cambridge, MA: MIT Press, 1997).

Churchland, P. S., "Consciousness: The transmutation of a concept," *Pacific Philosophical Quarterly* 64 (1983): pp. 80–93.

Crick, F., and Koch, C. "Towards a neurobiological theory of consciousness," *Seminars in the Neurosciences* 2 (1990): pp. 263–275.

Crick, F. *The Astonishing Hypothesis* (New York: Scribners. 1994).

Davies, M., and Humphreys, G. *Consciousness* (Oxford: Blackwell, 1993a).

_____. (1993b) Introduction. In Davies and Humphreys (1993a).

Dennett, D. *Consciousness Explained* (Boston: Little, Brown, 1991).

_____ "The message is: There is no medium," *Philosophy and Phenomenological Research* III (1993): p. 4.

_____ "Are we explaining consciousness yet?" *Cognition* 79 (2001): pp. 1–2, 221–37.

Dennett, D., and Kinsbourne, M., "Time and the observer: The where and when of consciousness in the brain," *Behavioral and Brain Sciences* 15 (1992a): pp. 183–200.

_____ "Escape from the Cartesian theater," *Behavioral and Brain Sciences* 15 (1992b): pp. 234–48.

Farah, M., "Visual perception and visual awareness after brain damage: A tutorial overview," in *Consciousness and Unconscious Information Processing: Attention and Performance*, Umilta and Moscovitch (Cambridge, MA: MIT Press., 1994).

Flanagan, O. *Consciousness Reconsidered* (Cambridge, MA: MIT Press, 1992).

Gallup, G., "Self-awareness and the emergence of mind in primates," *American Journal of Primatology* 2 (1982): pp. 237–48.

Goldman, A., "The psychology of folk psychology," *Behavioral and Brain Sciences* 16 no. 1 (1993): pp. 15–28.

_____ "Consciousness, folk psychology and cognitive science, *Consciousness and Cognition II* (1993b): p. 3.

Goodale, M., and Milner, D., "Separate visual pathways for perception and action," *Trends in Neuro-science* 15 (1992): pp. 20–25.

Hauser, M. D., Kralik, J., Botto, C., Garrett, M., and Oser, J., "Self-recognition in primates: Phylogeny and the salience of species-typical traits," *Proc. Nat. Acad. Sci.* 92 (1995): pp. 10811–14.

Hill, C. *Sensations; A Defense of Type Materialism* (New York: Cambridge University Press 1991).

Humphrey, N. *A History of the Mind* (New York: Simon & Schuster 1992).

Huxley, T. H. *Lessons in Elementary Psychology* 8, 1866 p. 210. Quoted in Humphrey, 1992.

Jackendoff, R. *Consciousness and the Computational Mind* (Cambridge, MA: MIT Press, 1987).

Jackson, F., "What Mary didn't know," *Journal of Philosophy* 83 (1986): pp. 291–95.

Kirk, R., "*Consciousness and concepts*," *Proceedings of the Aristotelian Society,* Supplementary Volume LXVI (1992): pp. 23–40.

Kuhn, T., "A function for thought experiments," in *Melanges Alexandre Koyre*, vol. 1 (Paris: Hermann, 1964), pp. 307–34.

Levine, J., "Review of Owen Flanagan's *Consciousness Reconsidered*," *Philosophical Review* 103 (1994): pp. 353–56.

Loar, B., "Phenomenal properties," in *Philosophical Perspectives: Action Theory and Philosophy of Mind,* J. Tomberlin, ed., (Atascadero, CA: Ridgeview, 1990).

Lormand, E., 1995. What qualitative consciousness is like. Manuscript.

Lycan, W. *Consciousness* (Cambridge, MA: MIT Press 1987).

McGinn, C. *The Problem of Consciousness* (Oxford, UK: Blackwell, 1991).

Malcolm, N., "Consciousness and causality," in D. M. Armstrong and N. Malcolm, *Consciousness and Causality* (Oxford, UK: Blackwell, 1984).

Marcel, A. J., "Consciousness and processing: Choosing and testing a null hypothesis," *The Behavioral and Brain Sciences* 9 (1986): pp. 40–41.

Nagel, T., "What is it like to be a bat?" *Philosophical Review* 83 (1974): pp. 435–50.

_____ *Mortal Questions* (Cambridege, UK: Cambridge University Press, 1979).

Natsoulas, T., "What is wrong with the appendage theory of consciousness?" *Philosophical Psychology* VI, no. 2 (1993): pp. 137–54.

Nelkin, N., "The connection between intentionality and consciousness," in Davies and Humphreys, *Consciousness* (1993a).

Povinelli, D., "What chimpanzees know about the mind," in *Behavioral Diversity in Chimpanzees* (Cambridge, MA: Harvard University Press, 1994).

Rey, G., "A reason for doubting the existence of consciousness," in *Consciousness and Self-Regulation*, vol 3, R. Davidson et al., eds., (New York: Plenum, 1983).

_____ "A question about consciousness," in *Perspectives on Mind*, H. Otto and J. Tuedio, eds., (Dordrecht, Holland: Reidel, 1988).

Rosenthal, D., "Two concepts of consciousness," *Philosophical Studies* 49 (1986): pp. 329–59.

_____ "A theory of consciousness," in *The Nature of Consciousness: Philosophical Debates,* N. Block et al., eds., (Cambridge, MA: MIT Press, 1997).

Schacter, D., "On the relation between memory and consciousness: Dissociable interactions and conscious experience," in *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving,* H. Roediger and F. Craik, eds., (Hillsdale, NJ: Erlbaum, 1989).

Searle, J. *Intentionality* (New York: Cambridge, 1983).

_____ "Who is computing with the brain?" *Behavioral and Brain Sciences* 13, no. 4 (1990): pp. 632–42.

_____ *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992).

Shoemaker, S., "Functionalism and qualia," *Philosophical Studies* 27 (1975): pp. 291–315.

_____ "The inverted spectrum." *The Journal of Philosophy* 74, no. 7 (1981): pp. 357–81.

Stich, S., "Autonomous psychology and the belief-desire thesis," *The Monist* 61 (1978): pp. 573–91.

Van Gulick, R., "What difference does consciousness make?" *Philosophical Topics* 17, no. 1 (1989): pp. 211–30.

_____ "Understanding the phenomenal mind: Are we all just armadillos?" in Davies and Humphreys, *Consciousness* (1993a).

Weiskrantz, L., "Some contributions of neuropsychology of vision and memory to the problem of consciousness," in Marcel and Bisiach, *Consciousness in Cocntemporary Science* (Oxford: Oxford University Press, 1988).

White, S. L., "What is it like to be an homunculus?" *Pacific Philosophical Quarterly* 68 (1987): pp. 148–74.

Wiser, M., and Carey, S., "When heat and temperature were one," in *Mental Models,* D. Gentner and A. Stevens, eds., (Hillsdale, NJ: Lawrence Erlbaum, 1983).

NOTES

1. Abridged (with changes by the author). I have changed only what seems mistaken even from the point of view of my former position. No attempt has been made to systematically update the references.
2. See Kuhn 1964 on velocity, and Block and Dworkin 1974 and Wiser and Covey 1983 on heat.
3. But what is it about thoughts that makes them P-conscious? One possibility is that it is just a series of mental images or sub vocalizations that make thoughts P-conscious. Another possibility is that the contents themselves have a P-conscious aspect independently of their vehicles. See Lormand 1995. and Burge 1997.
4. My view is that although P-conscious content cannot be *reduced to* or identified with intentional content (at least not on relatively a priori grounds), P-conscious contents often—maybe always—have an intentional aspect, representing in a primitive non-intentional way.
5. I know some will think that I invoked inverted and absent qualia a few paragraphs above when I described the explanatory gap as involving the question of why a creature with a brain which has a physiological and functional nature like ours couldn't have different experience or none at all. But the spirit of the question as I asked it allows for an answer that explains why such creatures cannot exist, and thus there is no presupposition that these are real possibilities.
6. I have been using the P-consciousness/A-consciousness distinction in my lectures for many years, but it only found its way into print in my 'Consciousness and Accessibility' (1990), and my 1991, 1992, 1993. My claims about the distinction have been criticized in Searle 1990, 1992 and Flanagan 1992—I reply to Flanagan below; and there is an illuminating discussion in Davies and Humphreys 1993b, a point of which will be taken up in a footnote to follow. See also Levine's (1994) review of Flanagan which discusses Flanagan's critique of the distinction. See also Kirk 1992 for an identification of P-consciousness with something like A-consciousness.
7. The full definition was: A state is access-conscious if, in virtue of one's having the state, a representation of its content is (1) inferentially promiscuous, that is, poised for use as a premise in reasoning, (2) poised for rational control of action, and (3) poised for rational control of speech.
8. Dennett 1991 and Dennett and Kinsbourne (1992) advocate the 'multiple drafts' account of consciousness. Dennett switched to the cerebral celebrity view in his 1993 paper.
9. See Dennett 2001 and Block 2001 for a more sophisticated treatment of this dialectic.
10. Some may say that only fully conceptualized content can play a role in reasoning, be reportable, and rationally control action. Such a view should not be adopted in isolation from views about which contents are personal and which are sub-personal.
11. The concept of P-consciousness is not a functional concept, however, I acknowledge the empirical possibility that the scientific nature of P-consciousness has something to do with information processing. We can ill afford to close off empirical possibilities given the difficulty of solving the mystery of P-consciousness.
12. The distinction has some similarity to the sensation/perception distinction; I won't take the space to lay out the differences. See Humphrey 1992 for an interesting discussion of the latter distinction.
13. If you are tempted to deny the existence of these states of the perceptual system, you should think back to the total zombie just mentioned. Putting aside the issue of the possibility of this zombie, note that on a computational notion of cognition, the zombie has *all* the same A-conscious contents that you have (if he is your computational duplicate). A-consciousness is an informational notion. The states of the superblindsighter's perceptual system are A-conscious for the same reason as the zombie's.
14. Farah claims that blindsight is more degraded than sight. But Weiskrantz 1988 notes that his patient DB had better acuity in some areas of the blind field (in some circumstances) than in his sighted field. It would be better to understand her 'degraded' in terms of lack of access.
15. Thus, there is a conflict between this physiological claim and the Schacter model which dictates that destroying the P-consciousness module will prevent A-consciousness.
16. See White 1987 for an account of why self-consciousness should be firmly distinguished from P-consciousness, and why self-consciousness is more relevant to certain issues of value.
17. The pioneer of these ideas in the philosophical literature is David Armstrong 1968, 1980. William Lycan 1987 has energetically pursued self-scanning, and David Rosenthal 1986, 1993; Peter Carruthers 1989, 1992; and Norton Nelkin 1993 have championed higher order thought. See also Natsoulas 1993. Lormand 1995 makes some powerful criticisms of Rosenthal.
18. To be fair to Rey, his argument is more like a dilemma: for any supposed feature of consciousness, either a laptop of the sort we have today has it or else you can't be sure you have it yourself. In the case of P-consciousness, laptops don't have it, and we are sure we do, so once we make these distinctions, his argument loses plausibility.
19. This represents a change of view from Block 1994, wherein I said that 'consciousness' ought to be ambiguous rather than saying it is now ambiguous.
20. I would like to thank Tyler Burge, Susan Carey, David Chalmers, Martin Davies, Wayne Davis, Bert Dreyfus, Guven Guzeldere, Paul Horwich, Jerry Katz, Leonard Katz, Joe Levine, David Rosenthal, Jerome Schaffer, Sydney Shoemaker, Stephen White, and Andrew Young for their very helpful comments on earlier versions of this paper. I have been giving this paper at colloquia and meetings since the fall of 1990, and I am grateful to the many audiences which have made interesting and useful comments, especially the audience at the conference on my work at the University of Barcelona in June 1993.

What Is It Like to Be a Bat?

Thomas Nagel

Consciousness is what makes the mind-body problem really intractable. Perhaps that is why current discussions of the problem give it little attention or get it obviously wrong. The recent wave of reductionist euphoria has produced several analyses of mental phenomena and mental concepts designed to explain the possibility of some variety of materialism, psychophysical identification, or reduction.[1] But the problems dealt with are those common to this type of reduction and other types, and what makes the mind-body problem unique, and unlike the water-H_2O problem or the Turing machine-IBM machine problem or the lightning-electrical discharge problem or the gene-DNA problem or the oak tree-hydrocarbon problem, is ignored.

Every reductionist has his favorite analogy from modern science. It is most unlikely that any of these unrelated examples of successful reduction will shed light on the relation of mind to brain. But philosophers share the general human weakness for explanations of what is incomprehensible in terms suited for what is familiar and well understood, though entirely different. This has led to the acceptance of implausible accounts of the mental largely because they would permit familiar kinds of reduction. I shall try to explain why the usual examples do not help us to understand the relation between mind and body—why, indeed, we have at present no conception of what an explanation of the physical nature of a mental phenomenon would be. Without consciousness the mind-body problem would be much less interesting. With consciousness it seems hopeless. The most important and characteristic feature of conscious mental phenomena is very poorly understood. Most reductionist theories do not even try to explain it. And careful examination will show that no currently available concept of reduction is applicable to it. Perhaps a new theoretical form can be devised for the purpose, but such a solution, if it exists, lies in the distant intellectual future.

Conscious experience is a widespread phenomenon. It occurs at many levels of animal life, though we cannot be sure of its presence in the simpler organisms, and it is very difficult to say in general what provides evidence of it. (Some extremists have been prepared to deny it even of mammals other than man.) No doubt it occurs in countless forms totally unimaginable to us, on other planets in other solar systems throughout the universe. But no matter how the form may vary, the fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism. There may be further implications about the form of the experience; there may even (though I doubt it) be implications about the behavior of the organism. But fundamentally an organism has conscious mental states if and only if there is something that it is like to *be* that organism—something it is like *for* the organism.

We may call this the subjective character of experience. It is not captured by any of the familiar, recently devised reductive analyses of the mental, for all of them are logically compatible with its absence. It is not analyzable in terms of any explanatory system of functional states, or intentional states, since these could be ascribed to robots or automata that behaved like people though they experienced nothing.[2] It is not analyzable in terms of the causal role of experiences in relation to typical human behavior—for similar reasons.[3] I do not deny that conscious mental states and events cause behavior, nor that they may be given functional characterizations. I deny only that this kind of thing exhausts their analysis. Any reductionist program has to be based on an analysis of what is to be reduced. If the analysis leaves something out, the problem will be falsely posed. It is useless to base the defense of materialism on any analysis of mental phenomena that fails to deal explicitly with their subjective character. For there is no reason to suppose that a reduction which seems plausible when no attempt is made to account for consciousness can be extended to include consciousness. Without some idea, therefore, of what the subjective character of experience is, we cannot know what is required of a physicalist theory.

From *Philosophical Review* 83 (1974): pp. 435–50.

While an account of the physical basis of mind must explain many things, this appears to be the most difficult. It is impossible to exclude the phenomenological features of experience from a reduction in the same way that one excludes the phenomenal features of an ordinary substance from a physical or chemical reduction of it—namely, by explaining them as effects on the minds of human observers.[4] If physicalism is to be defended, the phenomenological features must themselves be given a physical account. But when we examine their subjective character it seems that such a result is impossible. The reason is that every subjective phenomenon is essentially connected with a single point of view, and it seems inevitable that an objective, physical theory will abandon that point of view.

Let me first try to state the issue somewhat more fully than by referring to the relation between the subjective and the objective, or between the *pour-soi* and the *en-soi*. This is far from easy. Facts about what it is like to be an *X* are very peculiar, so peculiar that some may be inclined to doubt their reality, or the significance of claims about them. To illustrate the connection between subjectivity and a point of view, and to make evident the importance of subjective features, it will help to explore the matter in relation to an example that brings out clearly the divergence between the two types of conception, subjective and objective.

I assume we all believe that bats have experience. After all, they are mammals, and there is no more doubt that they have experience than that mice or pigeons or whales have experience. I have chosen bats instead of wasps or flounders because if one travels too far down the phylogenetic tree, people gradually shed their faith that there is experience there at all. Bats, although more closely related to us than those other species, nevertheless present a range of activity and a sensory apparatus so different from ours that the problem I want to pose is exceptionally vivid (though it certainly could be raised with other species). Even without the benefit of philosophical reflection, anyone who has spent some time in an enclosed space with an excited bat knows what it is to encounter a fundamentally *alien* form of life.

I have said that the essence of the belief that bats have experience is that there is something that it is like to be a bat. Now we know that most bats (the microchiroptera, to be precise) perceive the external world primarily by sonar, or echolocation, detecting the reflections, from objects within range, of their own rapid, subtly modulated, high-frequency shrieks. Their brains are designed to correlate the outgoing impulses with the subsequent echoes, and the information thus acquired enables bats to make precise discriminations of distance, size, shape, motion, and texture comparable to those we make by vision. But bat sonar, though clearly a form of perception, is not similar in its operation to any sense that we possess, and there is no reason to suppose that it is subjectively like anything we can experience or imagine. This appears to create difficulties for the notion of what it is like to be a bat. We must consider whether any method will permit us to extrapolate to the inner life of the bat from our own case,[5] and if not, what alternative methods there may be for understanding the notion.

Our own experience provides the basic material for our imagination, whose range is therefore limited. It will not help to try to imagine that one has webbing on one's arms, which enables one to fly around at dusk and dawn catching insects in one's mouth; that one has very poor vision, and perceives the surrounding world by a system of reflected high-frequency sound signals; and that one spends the day hanging upside down by one's feet in an attic. In so far as I can imagine this (which is not very far), it tells me only what it would be like for *me* to behave as a bat behaves. But that is not the question. I want to know what it is like for a *bat* to be a bat. Yet if I try to imagine this, I am restricted to the resources of my own mind, and those resources are inadequate to the task. I cannot perform it either by imagining additions to my present experience, or by imagining segments gradually subtracted from it, or by imagining some combination of additions, subtractions, and modifications.

To the extent that I could look and behave like a wasp or a bat without changing my fundamental structure, my experiences would not be anything like the experiences of those animals. On the other hand, it is doubtful that any meaning can be attached to the supposition that I should possess the internal neurophysiological constitution of a bat. Even if I could by gradual degrees be transformed into a bat, nothing in my present constitution enables me to imagine what the experiences of such a future stage of myself thus metamorphosed would be like. The best evidence would come from the experiences of bats, if we only knew what they were like.

So if extrapolation from our own case is involved in the idea of what it is like to be a bat,

the extrapolation must be incompletable. We cannot form more than a schematic conception of what it *is* like. For example, we may ascribe general *types* of experience on the basis of the animal's structure and behavior. Thus we describe bat sonar as a form of three-dimensional forward perception; we believe that bats feel some versions of pain, fear, hunger, and lust, and that they have other, more familiar types of perception besides sonar. But we believe that these experiences also have in each case a specific subjective character, which it is beyond our ability to conceive. And if there is conscious life elsewhere in the universe, it is likely that some of it will not be describable even in the most general experiential terms available to us.[6] (The problem is not confined to exotic cases, however, for it exists between one person and another. The subjective character of the experience of a person deaf and blind from birth is not accessible to me, for example, nor presumably is mine to him. This does not prevent us each from believing that the other's experience has such a subjective character.)

If anyone is inclined to deny that we can believe in the existence of facts like this whose exact nature we cannot possibly conceive, he should reflect that in contemplating the bats we are in much the same position that intelligent bats or Martians[7] would occupy if they tried to form a conception of what it was like to be us. The structure of their own minds might make it impossible for them to succeed, but we know they would be wrong to conclude that there is not anything precise that it is like to be us: that only certain general types of mental state could be ascribed to us (perhaps perception and appetite would be concepts common to us both; perhaps not). We know they would be wrong to draw such a skeptical conclusion because we know what it is like to be us. And we know that while it includes an enormous amount of variation and complexity, and while we do not possess the vocabulary to describe it adequately, its subjective character is highly specific, and in some respects describable in terms that can be understood only by creatures like us. The fact that we cannot expect ever to accommodate in our language a detailed description of Martian or bat phenomenology should not lead us to dismiss as meaningless the claim that bats and Martians have experiences fully comparable in richness of detail to our own. It would be fine if someone were to develop concepts and a theory that enabled us to think about those things; but such an understanding may be permanently denied to us by the limits of our nature. And to deny the reality or logical significance of what we can never describe or understand is the crudest form of cognitive dissonance.

This brings us to the edge of a topic that requires much more discussion than I can give it here: namely, the relation between facts on the one hand and conceptual schemes or systems of representation on the other. My realism about the subjective domain in all its forms implies a belief in the existence of facts beyond the reach of human concepts. Certainly it is possible for a human being to believe that there are facts which humans never *will* possess the requisite concepts to represent or comprehend. Indeed, it would be foolish to doubt this, given the finiteness of humanity's expectations. After all, there would have been transfinite numbers even if everyone had been wiped out by the Black Death before Cantor discovered them. But one might also believe that there are facts which *could* not ever be represented or comprehended by human beings, even if the species lasted forever—simply because our structure does not permit us to operate with concepts of the requisite type. This impossibility might even be observed by other beings, but it is not clear that the existence of such beings, or the possibility of their existence, is a precondition of the significance of the hypothesis that there are humanly inaccessible facts. (After all, the nature of beings with access to humanly inaccessible facts is presumably itself a humanly inaccessible fact.) Reflection on what it is like to be a bat seems to lead us, therefore, to the conclusion that there are facts that do not consist in the truth of propositions expressible in a human language. We can be compelled to recognize the existence of such facts without being able to state or comprehend them.

I shall not pursue this subject, however. Its bearing on the topic before us (namely, the mind-body problem) is that it enables us to make a general observation about the subjective character of experience. Whatever may be the status of facts about what it is like to be a human being, or a bat, or a Martian, these appear to be facts that embody a particular point of view.

I am not adverting here to the alleged privacy of experience to its possessor. The point of view in question is not one accessible only to a single individual. Rather it is a *type*. It is often possible to take up a point of view other than one's own, so the comprehension of such facts is not limited to one's own case. There is

a sense in which phenomenological facts are perfectly objective: one person can know or say of another what the quality of the other's experience is. They are subjective, however, in the sense that even this objective ascription of experience is possible only for someone sufficiently similar to the object of ascription to be able to adopt his point of view—to understand the ascription in the first person as well as in the third, so to speak. The more different from oneself the other experiencer is, the less success one can expect with this enterprise. In our own case we occupy the relevant point of view, but we will have as much difficulty understanding our own experience properly if we approach it from another point of view as we would if we tried to understand the experience of another species without taking up *its* point of view.[8]

This bears directly on the mind-body problem. For if the facts of experience—facts about what it is like *for* the experiencing organism—are accessible only from one point of view, then it is a mystery how the true character of experiences could be revealed in the physical operation of that organism. The latter is a domain of objective facts *par excellence*—the kind that can be observed and understood from many points of view and by individuals with differing perceptual systems. There are no comparable imaginative obstacles to the acquisition of knowledge about bat neurophysiology by human scientists, and intelligent bats or Martians might learn more about the human brain than we ever will.

This is not by itself an argument against reduction. A Martian scientist with no understanding of visual perception could understand the rainbow, or lightning, or clouds as physical phenomena, though he would never be able to understand the human concepts of rainbow, lightning, or cloud, or the place these things occupy in our phenomenal world. The objective nature of the things picked out by these concepts could be apprehended by him because, although the concepts themselves are connected with a particular point of view and a particular visual phenomenology, the things apprehended from that point of view are not: they are observable from the point of view but external to it; hence they can be comprehended from other points of view also, either by the same organisms or by others. Lightning has an objective character that is not exhausted by its visual appearance, and this can be investigated by a Martian without vision. To be precise, it has a *more* objective character than is revealed in its visual appearance. In speaking of the move from subjective to objective characterization, I wish to remain noncommittal about the existence of an end point, the completely objective intrinsic nature of the thing, which one might or might not be able to reach. It may be more accurate to think of objectivity as a direction in which the understanding can travel. And in understanding a phenomenon like lightning, it is legitimate to go as far away as one can from a strictly human viewpoint.[9]

In the case of experience, on the other hand, the connection with a particular point of view seems much closer. It is difficult to understand what could be meant by the *objective* character of an experience, apart from the particular point of view from which its subject apprehends it. After all, what would be left of what it was like to be a bat if one removed the viewpoint of the bat? But if experience does not have, in addition to its subjective character, an objective nature that can be apprehended from many different points of view, then how can it be supposed that a Martian investigating my brain might be observing physical processes which were my mental processes (as he might observe physical processes which were bolts of lightning), only from a different point of view? How, for that matter, could a human physiologist observe them from another point of view?[10]

We appear to be faced with a general difficulty about psychophysical reduction. In other areas the process of reduction is a move in the direction of greater objectivity, toward a more accurate view of the real nature of things. This is accomplished by reducing our dependence on individual or species-specific points of view toward the object of investigation. We describe it not in terms of the impressions it makes on our senses, but in terms of its more general effects and of properties detectable by means other than the human senses. The less it depends on a specifically human viewpoint, the more objective is our description. It is possible to follow this path because although the concepts and ideas we employ in thinking about the external world are initially applied from a point of view that involves our perceptual apparatus, they are used by us to refer to things beyond themselves—toward which we *have* the phenomenal point of view. Therefore we can abandon it in favor of another, and still be thinking about the same things.

Experience itself, however, does not seem to fit the pattern. The idea of moving from appearance to reality seems to make no sense here.

What is the analogue in this case to pursuing a more objective understanding of the same phenomena by abandoning the initial subjective viewpoint toward them in favor of another that is more objective but concerns the same thing? Certainly it *appears* unlikely that we will get closer to the real nature of human experience by leaving behind the particularity of our human point of view and striving for a description in terms accessible to beings that could not imagine what it was like to be us. If the subjective character of experience is fully comprehensible only from one point of view, then any shift to greater objectivity—that is, less attachment to a specific viewpoint—does not take us nearer to the real nature of the phenomenon: it takes us farther away from it.

In a sense, the seeds of this objection to the reducibility of experience are already detectable in successful cases of reduction; for in discovering sound to be, in reality, a wave phenomenon in air or other media, we leave behind one viewpoint to take up another, and the auditory, human or animal viewpoint that we leave behind remains unreduced. Members of radically different species may both understand the same physical events in objective terms, and this does not require that they understand the phenomenal forms in which those events appear to the senses of members of the other species. Thus it is a condition of their referring to a common reality that their more particular viewpoints are not part of the common reality that they both apprehend. The reduction can succeed only if the species-specific viewpoint is omitted from what is to be reduced.

But while we are right to leave this point of view aside in seeking a fuller understanding of the external world, we cannot ignore it permanently, since it is the essence of the internal world, and not merely a point of view on it. Most of the neobehaviorism of recent philosophical psychology results from the effort to substitute an objective concept of mind for the real thing, in order to have nothing left over which cannot be reduced. If we acknowledge that a physical theory of mind must account for the subjective character of experience, we must admit that no presently available conception gives us a clue how this could be done. The problem is unique. If mental processes are indeed physical processes, then there is something it is like, intrinsically,[11] to undergo certain physical processes. What it is for such a thing to be the case remains a mystery.

What moral should be drawn from these reflections, and what should be done next? It would be a mistake to conclude that physicalism must be false. Nothing is proved by the inadequacy of physicalist hypotheses that assume a faulty objective analysis of mind. It would be truer to say that physicalism is a position we cannot understand because we do not at present have any conception of how it might be true. Perhaps it will be thought unreasonable to require such a conception as a condition of understanding. After all, it might be said, the meaning of physicalism is clear enough: mental states are states of the body; mental events are physical events. We do not know *which* physical states and events they are, but that should not prevent us from understanding the hypothesis. What could be clearer than the words 'is' and 'are'?

But I believe it is precisely this apparent clarity of the word 'is' that is deceptive. Usually, when we are told that *X* is *Y* we know *how* it is supposed to be true, but that depends on a conceptual or theoretical background and is not conveyed by the 'is' alone. We know how both '*X*' and '*Y*' refer, and the kinds of things to which they refer, and we have a rough idea how the two referential paths might converge on a single thing, be it an object, a person, a process, an event, or whatever. But when the two terms of the identification are very disparate it may not be so clear how it could be true. We may not have even a rough idea of how the two referential paths could converge, or what kind of things they might converge on, and a theoretical framework may have to be supplied to enable us to understand this. Without the framework, an air of mysticism surrounds the identification.

This explains the magical flavor of popular presentations of fundamental scientific discoveries, given out as propositions to which one must subscribe without really understanding them. For example, people are now told at an early age that all matter is really energy. But despite the fact that they know what 'is' means, most of them never form a conception of what makes this claim true, because they lack the theoretical background.

At the present time the status of physicalism is similar to that which the hypothesis that matter is energy would have had if uttered by a pre-Socratic philosopher. We do not have the beginnings of a conception of how it might be true. In order to understand the hypothesis that a mental event is a physical event, we require more than an understanding of the word 'is.'

The idea of how a mental and a physical term might refer to the same thing is lacking, and the usual analogies with theoretical identification in other fields fail to supply it. They fail because if we construe the reference of mental terms to physical events on the usual model, we either get a reappearance of separate subjective events as the effects through which mental reference to physical events is secured, or else we get a false account of how mental terms refer (for example, a causal behaviorist one).

Strangely enough, we may have evidence for the truth of something we cannot really understand. Suppose a caterpillar is locked in a sterile safe by someone unfamiliar with insect metamorphosis, and weeks later the safe is reopened, revealing a butterfly. If the person knows that the safe has been shut the whole time, he has reason to believe that the butterfly is or was once the caterpillar, without having any idea in what sense this might be so. (One possibility is that the caterpillar contained a tiny winged parasite that devoured it and grew into the butterfly.)

It is conceivable that we are in such a position with regard to physicalism. Donald Davidson has argued that if mental events have physical causes and effects, they must have physical descriptions. He holds that we have reason to believe this even though we do not—and in fact *could* not—have a general psychophysical theory.[12] His argument applies to intentional mental events, but I think we also have some reason to believe that sensations are physical processes, without being in a position to understand how. Davidson's position is that certain physical events have irreducibly mental properties, and perhaps some view describable in this way is correct. But nothing of which we can now form a conception corresponds to it; nor have we any idea what a theory would be like that enabled us to conceive of it.[13]

Very little work has been done on the basic question (from which mention of the brain can be entirely omitted) whether any sense can be made of experiences' having an objective character at all. Does it make sense, in other words, to ask what my experiences are *really* like, as opposed to how they appear to me? We cannot genuinely understand the hypothesis that their nature is captured in a physical description unless we understand the more fundamental idea that they *have* an objective nature (or that objective processes can have a subjective nature).[14]

I should like to close with a speculative proposal. It may be possible to approach the gap between subjective and objective from another direction. Setting aside temporarily the relation between the mind and the brain, we can pursue a more objective understanding of the mental in its own right. At present we are completely unequipped to think about the subjective character of experience without relying on the imagination—without taking up the point of view of the experiential subject. This should be regarded as a challenge to form new concepts and devise a new method—an objective phenomenology not dependent on empathy or the imagination. Though presumably it would not capture everything, its goal would be to describe, at least in part, the subjective character of experiences in a form comprehensible to beings incapable of having those experiences.

We would have to develop such a phenomenology to describe the sonar experiences of bats; but it would also be possible to begin with humans. One might try, for example, to develop concepts that could be used to explain to a person blind from birth what it was like to see. One would reach a blank wall eventually, but it should be possible to devise a method of expressing in objective terms much more than we can at present, and with much greater precision. The loose intermodal analogies—for example, 'Red is like the sound of a trumpet'—which crop up in discussions of this subject are of little use. That should be clear to anyone who has both heard a trumpet and seen red. But structural features of perception might be more accessible to objective description, even though something would be left out. And concepts alternative to those we learn in the first person may enable us to arrive at a kind of understanding even of our own experience which is denied us by the very ease of description and lack of distance that subjective concepts afford.

Apart from its own interest, a phenomenology that is in this sense objective may permit questions about the physical[15] basis of experience to assume a more intelligible form. Aspects of subjective experience that admitted this kind of objective description might be better candidates for objective explanations of a more familiar sort. But whether or not this guess is correct, it seems unlikely that any physical theory of mind can be contemplated until more thought has been given to the general problem of subjective and objective. Otherwise we cannot even pose the mind-body problem without sidestepping it.[16]

NOTES

1. Examples are J. J. C. Smart, *Philosophy and Scientific Realism* (London: Routledge, 1963); David K. Lewis, "An Argument for the Identity Theory," *Journal of Philosophy*, LXIII (1966), reprinted with addenda in David M. Rosenthal, *Materialism & the Mind-Body Problem* (Englewood Cliffs, NJ: Prentice Hall, 1971); Hilary Putnam, "Psychological Predicates" in W. H. Capitan and D. D. Merrill, *Art, Mind, & Religion* (Pittsburgh: University of Pittsburgh Press, 1967), reprinted in Rosenthal, op. cit., as "The Nature of Mental States"; D. M. Armstrong, *A Materialist Theory of the Mind* (London: Routledge & Kegan Paul, 1968); D. C. Dennett, *Content and Consciousness* (London: Routledge & Kegan Paul, 1969). I have expressed earlier doubts in "Armstrong on the Mind," *Philosophical Review* LXXIX (1970): pp. 394–403; "Brain Bisection and the Unity of Consciousness," *Synthèse* 22 (1971); and a review of Dennett, *Journal of Philosophy* LXIX (1972). See also Saul Kripke, "Naming and Necessity" in D. Davidson and G. Harman, *Semantics of Natural Language* (Dordrecht: D. Reidel, 1972), especially pp. 334–42; and M. T. Thornton, "Ostensive Terms and Materialism," *The Monist* 56 (1972).
2. Perhaps there could not actually be such robots. Perhaps anything complex enough to behave like a person would have experiences. But that, if true, is a fact which cannot be discovered merely by analyzing the concept of experience.
3. It is not equivalent to that about which we are incorrigible, both because we are not incorrigible about experience and because experience is present in animals lacking language and thought, who have no beliefs at all about their experiences.
4. Cf. Richard Rorty, "Mind-Body Identity, Privacy, and Categories," *The Review of Metaphysics*, XIX (1965), especially 37–38.
5. By 'our own case' I do not mean just 'my own case,' but rather the mentalistic ideas that we apply unproblematically to ourselves and other human beings.
6. Therefore the analogical form of the English expression 'what it is *like*' is misleading. It does not mean 'what (in our experience) it *resembles*,' but rather 'how it is for the subject himself.'
7. Any intelligent extraterrestrial beings totally different from us.
8. It may be easier than I suppose to transcend interspecies barriers with the aid of the imagination. For example, blind people are able to detect objects near them by a form of sonar, using vocal clicks or taps of a cane. Perhaps if one knew what that was like, one could by extension imagine roughly what it was like to possess the much more refined sonar of a bat. The distance between oneself and other persons and other species can fall anywhere on a continuum. Even for other persons the understanding of what it is like to be them is only partial, and when one moves to species very different from oneself, a lesser degree of partial understanding may still be available. The imagination is remarkably flexible. My point, however, is not that we cannot *know* what it is like to be a bat. I am not raising that epistemological problem. My point is rather that even to form a *conception* of what it is like to be a bat (and a fortiori to know what it is like to be a bat) one must take up the bat's point of view. If one can take it up roughly, or partially, then one's conception will also be rough or partial. Or so it seems in our present state of understanding.
9. The problem I am going to raise can therefore be posed even if the distinction between more subjective and more objective descriptions or viewpoints can itself be made only within a larger human point of view. I do not accept this kind of conceptual relativism, but it need not be refuted to make the point that psychophysical reduction cannot be accommodated by the subjective-to-objective model familiar from other cases.
10. The problem is not just that when I look at the 'Mona Lisa,' my visual experience has a certain quality, no trace of which is to be found by someone looking into my brain. For even if he did observe there a tiny image of the 'Mona Lisa,' he would have no reason to identify it with the experience.
11. The relation would therefore not be a contingent one, like that of a cause and its distinct effect. It would be necessarily true that a certain physical state felt a certain way. Saul Kripke (op. cit.) argues that causal behaviorist and related analyses of the mental fail because they construe, e.g., 'pain' as a merely contingent name of pains. The subjective character of an experience ('its immediate phenomenological quality' Kripke calls it [p. 340]) is the essential property left out by such analyses, and the one in virtue of which it is, necessarily, the experience it is. My view is closely related to his. Like Kripke, I find the hypothesis that a certain brain state should *necessarily* have a certain subjective character incomprehensible without further explanation. No such explanation emerges from theories which view the mind-brain relation as contingent, but perhaps there are other alternatives, not yet discovered.

    A theory that explained how the mind-brain relation was necessary would still leave us with Kripke's problem of explaining why it nevertheless appears contingent. That difficulty seems to me surmountable, in the following way. We may imagine something by representing it to ourselves either perceptually, sympathetically, or symbolically. I shall not try to say how symbolic imagination works, but part of what happens in the other two cases is this. To imagine something perceptually, we put ourselves in a conscious state resembling the state we would be in if we perceived it. To imagine something sympathetically, we put ourselves in a conscious state resembling the thing itself. (This method can be used only to imagine mental events and states—our own or another's.) When we try to imagine a mental state occurring without its associated brain state, we first sympathetically imagine the occurrence of the mental state: that is, we put ourselves into a state that resembles it mentally. At the same time, we attempt to perceptually imagine the non-occurrence of the associated physical state, by putting ourselves into another state unconnected with the first: one resembling that which we would be in if we perceived the non-occurrence of the physical state. Where the imagination of physical features is perceptual and the imagination of mental features is sympathetic, it appears to us that we can imagine any experience

occurring without its associated brain state, and vice versa. The relation between them will appear contingent even if it is necessary, because of the independence of the disparate types of imagination.

(Solipsism, incidentally, results if one misinterprets sympathetic imagination as if it worked like perceptual imagination: it then seems impossible to imagine any experience that is not one's own.)

12 See "Mental Events" in L. Foster and J.W. Swanson, *Experience and Theory* (Amherst, MA: University of Massachusetts Press, 1970); though I don't understand the argument against psychophysical laws.

13 Similar remarks apply to my paper "Physicalism," *Philosophical Review* LXXIV (1965): pp. 339–56, reprinted with postscript in John O'Connor, *Modern Materialism* (New York: Harcourt, Brace, and World, 1969).

14 This question also lies at the heart of the problem of other minds, whose close connection with the mind-body problem is often overlooked. If one understood how subjective experience could have an objective nature, one would understand the existence of subjects other than oneself.

15 I have not defined the term 'physical.' Obviously it does not apply just to what can be described by the concepts of contemporary physics, since we expect further developments. Some may think there is nothing to prevent mental phenomena from eventually being recognized as physical in their own right. But whatever else may be said of the physical, it has to be objective. So if our idea of the physical ever expands to include mental phenomena, it will have to assign them an objective character—whether or not this is done by analyzing them in terms of other phenomena already regarded as physical. It seems to me more likely, however, that mental-physical relations will eventually be expressed in a theory whose fundamental terms cannot be placed clearly in either category.

16 I have read versions of this paper to a number of audiences and am indebted to many people for their comments.

Quining Qualia

Daniel C. Dennett

1. Corralling the Quicksilver

'Qualia' is an unfamiliar term for something that could not be more familiar to each of us: the *ways things seem to us.* As is so often the case with philosophical jargon, it is easier to give examples than to give a definition of the term. Look at a glass of milk at sunset; *the way it looks to you*—the particular, personal, subjective visual quality of the glass of milk is the *quale* of your visual experience at the moment. The *way the milk tastes to you then* is another, gustatory, *quale*, and *how it sounds to you* as you swallow is an auditory *quale.* These various 'properties of conscious experience' are prime examples of *qualia.* Nothing, it seems, could you know more intimately than your own qualia; let the entire universe be some vast illusion, some mere figment of Descartes's evil demon, and yet what the figment is *made of* (for you) will be the *qualia* of your hallucinatory experiences. Descartes claimed to doubt everything that could be doubted, but he never doubted that his conscious experiences had qualia, the properties by which he knew or apprehended them.

The verb 'to quine' is even more esoteric. It comes from *The Philosophical Lexicon* (Dennett 1978c, 8th edition 1987), a satirical dictionary of eponyms: 'quine, *v.* To deny resolutely the existence or importance of something real or significant.' At first blush it would be hard to imagine a more quixotic quest than trying to convince people that there are no such properties as qualia; hence the ironic title of this chapter. But I am not kidding.

My goal is subversive. I am out to overthrow an idea that, in one form or another, is 'obvious' to most people—to scientists, philosophers, lay people. My quarry is frustratingly elusive; no sooner does it retreat in the face of one argument than 'it' reappears, apparently innocent of all charges, in a new guise.

Which idea of qualia am I trying to extirpate? Everything real has properties, and since I don't deny the reality of conscious experience, I grant that conscious experience

Reprinted from A. Marcel and E. Bisiach, eds., *Consciousness in Contemporary Science* (Oxford University Press, 1988), with permission of Oxford University Press.

has properties. I grant moreover that each person's states of consciousness have properties in virtue of which those states have the experiential content that they do. That is to say, whenever someone experiences something as being one way rather than another, this is true in virtue of some property of something happening in them at the time, but these properties are so unlike the properties traditionally imputed to consciousness that it would be grossly misleading to call any of them the long-sought qualia. Qualia are supposed to be *special* properties, in some hard-to-define way. My claim—which can only come into focus as we proceed—is that conscious experience has *no* properties that are special in *any* of the ways qualia have been supposed to be special.

The standard reaction to this claim is the complacent acknowledgment that while some people may indeed have succumbed to one confusion or fanaticism or another, one's own appeal to a modest, innocent notion of properties of subjective experience is surely safe. It is just that presumption of innocence I want to overthrow. I want to shift the burden of proof, so that anyone who wants to appeal to private, subjective properties has to prove first that in so doing they are *not* making a mistake. This status of *guilty until proven innocent* is neither unprecedented nor indefensible (so long as we restrict ourselves to concepts). Today, no biologist would dream of supposing that it was quite all right to appeal to some innocent concept of *élan vital.* Of course one *could* use the term to mean something in good standing; one could use *élan vital* as one's name for DNA, for instance, but this would be foolish nomenclature, considering the deserved suspicion with which the term is nowadays burdened. I want to make it just as uncomfortable for anyone to talk of qualia—or 'raw feels' or 'phenomenal properties' or 'subjective and intrinsic properties' or 'the qualitative character' of experience—with the standard presumption that they, and everyone else, knows what on earth they are talking about.[1]

What are qualia, *exactly?* This obstreperous query is dismissed by one author ('only half in jest') by invoking Louis Armstrong's legendary reply when asked what jazz was: 'If you got to ask, you ain't never gonna get to know.' (Block 1978, p. 281). This amusing tactic perfectly illustrates the presumption that is my target. If I succeed in my task, this move, which passes muster in most circles today, will look as quaint and insupportable as a jocular appeal to the ludicrousness of a living thing—a living thing, mind you!—doubting the existence of *élan vital.*

My claim, then, is not just that the various technical or theoretical concepts of qualia are vague or equivocal, but that the source concept, the 'pretheoretical' notion of which the former are presumed to be refinements, is so thoroughly confused that even if we undertook to salvage some 'lowest common denominator' from the theoreticians' proposals, any acceptable version would have to be so radically unlike the ill-formed notions that are commonly appealed to that it would be tactically obtuse—not to say Pickwickian—to cling to the term. Far better, tactically, to declare that there simply are no qualia at all.[2]

Rigorous arguments only work on well-defined materials, and since my goal is to destroy our faith in the pretheoretical or 'intuitive' concept, the right tools for my task are intuition pumps, not formal arguments. What follows is a series of fifteen intuition pumps, posed in a sequence designed to flush out—and then flush away—the offending intuitions. In section 2, I will use the first two intuition pumps to focus attention on the traditional notion. It will be the burden of the rest of the paper to convince you that these two pumps, for all their effectiveness, mislead us and should be discarded. In section 3, the next four intuition pumps create and refine a 'paradox' lurking in the tradition. This is not a formal paradox, but only a very powerful argument pitted against some almost irresistibly attractive ideas. In section 4, six more intuition pumps are arrayed in order to dissipate the attractiveness of those ideas, and section 5 drives this point home by showing how hapless those ideas prove to be when confronted with some real cases of anomalous experience. This will leave something of a vacuum, and in the final section three more intuition pumps are used to introduce and motivate some suitable replacements for the banished notions.

2. The Special Properties of Qualia

Intuition pump #1: watching you eat cauliflower. I see you tucking eagerly into a helping of steaming cauliflower, the merest whiff of which makes me faintly nauseated, and I find myself wondering how you could possibly relish *that taste*, and then it occurs to me that to you, cauliflower probably tastes (must taste?) different. A plausible hypothesis, it seems, especially since I

know that the very same food often tastes different to me at different times. For instance, my first sip of breakfast orange juice tastes much sweeter than my second sip if I interpose a bit of pancakes and maple syrup, but after a swallow or two of coffee, the orange juice goes back to tasting (roughly? exactly?) the way it did the first sip. Surely we want to say (or think about) such things, and surely we are not wildly wrong when we do, so . . . surely it is quite OK to talk of *the way the juice tastes to Dennett at time t*, and ask whether it is just the same as or different from *the way the juice tastes to Dennett at time t'* or *the way the juice tastes to Jones at time t*.

This 'conclusion' seems innocent, but right here we have already made the big mistake. The final step presumes that we can isolate the qualia from everything else that is going on—at least in principle or for the sake of argument. What counts as *the way the juice tastes to x* can be distinguished, one supposes, from what is a mere accompaniment, contributory cause, or by-product of this 'central' way. One dimly imagines taking such cases and stripping them down gradually to the essentials, leaving their common residuum, the way things look, sound, feel, taste, smell to various individuals at various times, independently of how those individuals are stimulated or non-perceptually affected, and independently of how they are subsequently disposed to behave or believe. The mistake is not in supposing that we can in practice ever or always perform this act of purification with certainty, but the more fundamental mistake of supposing that there is such a residual property to take seriously, however uncertain our actual attempts at isolation of instances might be.

The examples that seduce us are abundant in every modality. I cannot imagine, will never know, could never know, it seems, how Bach sounded to Glenn Gould. (I can barely recover in my memory the way Bach sounded to me when I was a child.) And I cannot know, it seems, what it is like to be a bat (Nagel 1974), or whether you see what I see, colorwise, when we look up at a clear 'blue' sky. The homely cases convince us of the reality of these special properties—those subjective tastes, looks, aromas, sounds—that we then apparently isolate for definition by this philosophical distillation.

The specialness of these properties is hard to pin down, but can be seen at work in *intuition pump #2: the wine-tasting machine.* Could Gallo Brothers replace their human wine tasters with a machine? A computer-based 'expert system' for quality control and classification is probably within the bounds of existing technology. We now know enough about the relevant chemistry to make the transducers that would replace taste buds and olfactory organs (delicate color vision would perhaps be more problematic), and we can imagine using the output of such transducers as the raw material—the 'sense data' in effect—for elaborate evaluations, descriptions, classifications. Pour the sample in the funnel and, in a few minutes or hours, the system would type out a chemical assay, along with commentary: 'a flamboyant and velvety Pinot, though lacking in stamina'—or words to such effect. Such a machine might well perform better than human wine tasters on all reasonable tests of accuracy and consistency the winemakers could devise,[3] but *surely* no matter how 'sensitive' and 'discriminating' such a system becomes, it will never have, and enjoy, what *we* do when we taste a wine: the qualia of conscious experience! Whatever informational, dispositional, functional properties its internal states have, none of them will be special in the way qualia are. If you share that intuition, you believe that there are qualia in the sense I am targeting for demolition.

What is special about qualia? Traditional analyses suggest some fascinating second-order properties of these properties. First, since one *cannot say* to another, no matter how eloquent one is and no matter how cooperative and imaginative one's audience is, exactly what way one is currently seeing, tasting, smelling and so forth, qualia are *ineffable*—in fact the paradigm cases of ineffable items. According to tradition, at least part of the reason why qualia are ineffable is that they are *intrinsic* properties—which seems to imply *inter alia* that they are somehow atomic and unanalyzable. Since they are 'simple' or 'homogeneous' there is nothing to get hold of when trying to describe such a property to one unacquainted with the particular instance in question.

Moreover, verbal comparisons are not the only cross-checks ruled out. *Any* objective, physiological or 'merely behavioral' test—such as those passed by the imaginary wine-tasting system—would of necessity miss the target (one can plausibly argue), so all interpersonal comparisons of these ways-of-appearing are (apparently) systematically impossible. In other words, qualia are essentially *private* properties. And, finally, since they *are* properties of *my experiences* (they're not chopped liver,

and they're not properties of, say, my cerebral blood flow—or haven't you been paying attention?), qualia are essentially directly accessible to the consciousness of their experiencer (whatever that means) or qualia are properties of one's experience with which one is intimately or directly acquainted (whatever that means) or 'immediate phenomenological qualities' (Block 1978) (whatever that means). They are, after all, the very properties the appreciation of which permits us to identify our conscious states. So, to summarize the tradition, qualia are supposed to be properties of a subject's mental states that are

1. ineffable
2. intrinsic
3. private
4. directly or immediately apprehensible in consciousness

Thus are qualia introduced onto the philosophical stage. They have seemed to be very significant properties to some theorists because they have seemed to provide an insurmountable and unavoidable stumbling block to functionalism, or more broadly, to materialism, or more broadly still, to any purely 'third-person' objective viewpoint or approach to the world (Nagel 1986). Theorists of the contrary persuasion have patiently and ingeniously knocked down all the arguments, and said most of the right things, but they have made a tactical error, I am claiming, of saying in one way or another: 'We theorists can handle *those qualia* you talk about just fine; we will show that you are just slightly in error about the nature of qualia.' What they ought to have said is: 'What qualia?'

My challenge strikes some theorists as outrageous or misguided because they think they have a much blander and hence less vulnerable notion of qualia to begin with. They think I am setting up and knocking down a strawman, and ask, in effect: 'Who said qualia are ineffable, intrinsic, private, directly apprehensible ways things seem to one?' Since my suggested fourfold essence of qualia may strike many readers as tendentious, it may be instructive to consider, briefly, an apparently milder alternative: qualia are simply 'the qualitative or phenomenal features of sense experience[s], in virtue of having which they resemble and differ from each other, qualitatively, in the ways they do' (Shoemaker 1982, p. 367). Surely I do not mean to deny *those* features!

I reply: it all depends on what 'qualitative or phenomenal' comes to. Shoemaker contrasts *qualitative* similarity and difference with 'intentional' similarity and difference—similarity and difference of the properties an experience represents or is 'of.' That is clear enough, but what then of 'phenomenal'? Among the non-intentional (and hence qualitative?) properties of my visual states are their physiological properties. Might these very properties be the qualia Shoemaker speaks of? It is supposed to be obvious, I take it, that these sorts of features are ruled out, because they are not 'accessible to introspection' (Shoemaker, private correspondence). These are features of my visual *state*, perhaps, but not of my visual *experience*. They are not *phenomenal* properties.

But then another non-intentional similarity some of my visual states share is that they tend to make me think about going to bed. I think this feature of them *is* accessible to introspection—on any ordinary, pretheoretical construal. Is that a phenomenal property or not? The term 'phenomenal' means nothing obvious and untendentious to me, and looks suspiciously like a gesture in the direction leading back to ineffable, private, directly apprehensible ways things seem to one.[4]

I suspect, in fact, that many are unwilling to take my radical challenge seriously largely because they want so much for qualia to be acknowledged. Qualia seem to many people to be the last ditch defense of the inwardness and elusiveness of our minds, a bulwark against creeping mechanism. They are sure there must be *some* sound path from the homely cases to the redoubtable category of the philosophers, since otherwise their last bastion of specialness will be stormed by science.

This special status for these presumed properties has a long and eminent tradition. I believe it was Einstein who once advised us that science could not give us the *taste* of the soup. Could such a wise man have been wrong? Yes, if he is taken to have been trying to remind us of the qualia that hide forever from objective science in the subjective inner sancta of our minds. There are no such things. Another wise man said so—Wittgenstein (1958, especially pp. 91–100). Actually, what he said was:

> The thing in the box has no place in the language-game at all; not even as a *something;* for the box might even be empty—No, one can 'divide through' by the thing in the box; it cancels out, whatever it is. (p. 100)

and then he went on to hedge his bets by saying 'It is not a *something*, but not a *nothing* either!

The conclusion was only that a nothing would serve just as well as a something about which nothing could be said' (p. 102). Both Einstein's and Wittgenstein's remarks are endlessly amenable to exegesis, but rather than undertaking to referee this War of the Titans, I choose to take what may well be a more radical stand than Wittgenstein's.[5] Qualia are not even 'something about which nothing can be said'; 'qualia' is a philosopher's term which fosters[6] nothing but confusion, and refers in the end to no properties or features at all.

3. The Traditional Paradox Regained

Qualia have not always been in good odor among philosophers. Although many have thought, along with Descartes and Locke, that it made sense to talk about private, ineffable properties of minds, others have argued that this is strictly nonsense—however naturally it trips off the tongue. It is worth recalling how qualia were presumably rehabilitated as properties to be taken seriously in the wake of Wittgensteinian and verificationist attacks on them as pseudo-hypotheses. The original version of *intuition pump #3: the inverted spectrum* (Locke 1690: II, xxxii, 15) is a speculation about two people: how do I know that you and I see the same subjective color when we look at something? Since we both learned color words by being shown public colored objects, our verbal behavior will match *even if we experience entirely different subjective colors.* The intuition that this hypothesis is systematically unconfirmable (and undisconfirmable, of course) has always been quite robust, but some people have always been tempted to think technology could (in principle) bridge the gap.

Suppose, in *intuition pump #4: the Brainstorm machine*, there were some neuroscientific apparatus that fits on your head and feeds your visual experience into my brain (as in the movie, *Brainstorm*, which is not to be confused with the book, *Brainstorms*). With eyes closed I accurately report everything you are looking at, except that I marvel at how the sky is yellow, the grass red, and so forth. Would this not confirm, empirically, that our qualia were different? But suppose the technician then pulls the plug on the connecting cable, inverts it 180 degrees and reinserts it in the socket. Now I report the sky is blue, the grass green, and so forth. Which is the 'right' orientation of the plug? Designing and building such a device would require that its 'fidelity' be tuned or calibrated by the normalization of the two subjects' reports—so we would be right back at our evidential starting point. The moral of this intuition pump is that no intersubjective comparison of qualia is possible, even with perfect technology.

So matters stood until someone dreamt up the presumably improved version of the thought experiment: the *intra*personal inverted spectrum. The idea seems to have occurred to several people independently (Gert 1965; Putnam 1965; Taylor 1966; Shoemaker 1969, 1975; Lycan 1973). Probably Block and Fodor 1972 have it in mind when they say 'It seems to us that the standard verificationist counterarguments against the view that the 'inverted spectrum' hypothesis is conceptually incoherent are not persuasive' (p. 172). In this version, *intuition pump #5: the neurosurgical prank*, the experiences to be compared are all in one mind. You wake up one morning to find that the grass has turned red, the sky yellow, and so forth. No one else notices any color anomalies in the world, so the problem must be in you. You are entitled, it seems, to conclude that you have undergone visual color qualia inversion (and we later discover, if you like, just how the evil neurophysiologists tampered with your neurons to accomplish this).

Here it seems at first—and indeed for quite a while—that qualia are acceptable properties after all, because propositions about them can be justifiably asserted, empirically verified and even explained. After all, in the imagined case, we can tell a tale in which we confirm a detailed neurophysiological account of the precise etiology of the dramatic change you undergo. It is tempting to suppose, then, that neurophysiological evidence, incorporated into a robust and ramifying theory, would have all the resolving power we could ever need for determining whether or not someone's qualia have actually shifted.

But this is a mistake. It will take some patient exploration to reveal the mistake in depth, but the conclusion can be reached—if not secured—quickly with the help of *intuition pump #6: alternative neurosurgery.* There are (at least) two different ways the evil neurosurgeon might create the inversion effect described in intuition pump #5:

1. Invert one of the 'early' qualia-producing channels, e.g. in the optic nerve, so that all relevant neural events 'downstream' are

the 'opposite' of their original and normal values. *Ex hypothesi* this inverts your qualia.

2. Leave all those early pathways intact and simply invert certain memory-access links—whatever it is that accomplishes your tacit (and even unconscious!) comparison of today's hues with those of yore. *Ex hypothesi* this does *not* invert your qualia at all, but just your memory-anchored dispositions to react to them.

On waking up and finding your visual world highly anomalous, you should exclaim 'Egad! *Something* has happened! Either my qualia have been inverted or my memory-linked qualia reactions have been inverted. I wonder which!'

The intrapersonal inverted spectrum thought experiment was widely supposed to be an improvement, since it moved the needed comparison into one subject's head. But now we can see that this is an illusion, since the link to earlier experiences, the link via memory, is analogous to the imaginary cable that might link two subjects in the original version.

This point is routinely—one might say traditionally—missed by the constructors of 'intra-subjective inverted spectrum' thought experiments, who suppose that the subject's *noticing the difference*—surely a vivid experience of discovery by the subject—would have to be an instance of (directly? incorrigibly?) recognizing the difference *as a shift in qualia.* But as my example shows, we could achieve the same startling effect in a subject without tampering with his presumed qualia at all. Since *ex hypothesi* the two different surgical invasions can produce exactly the same introspective effects while only one operation inverts the qualia, nothing in the subject's experience can favor one of the hypotheses over the other. So unless he seeks outside help, the state of his own qualia must be as unknowable to him as the state of anyone else's qualia. Hardly the privileged access or immediate acquaintance or direct apprehension the friends of qualia had supposed 'phenomenal features' to enjoy!

The outcome of this series of thought experiments is an intensification of the 'verificationist' argument against qualia. *If* there are qualia, they are even less accessible to our ken than we had thought. Not only are the classical intersubjective comparisons impossible (as the Brainstorm machine shows), but we cannot tell in our own cases whether our qualia have been inverted—at least not by introspection. It is surely tempting at this point—especially to non-philosophers—to decide that this paradoxical result must be an artifact of some philosophical misanalysis or other, the sort of thing that might well happen if you took a perfectly good pretheoretical notion—our everyday notion of qualia—and illicitly stretched it beyond the breaking point. The philosophers have made a mess; let them clean it up; meanwhile we others can get back to work, relying as always on our sober and unmetaphysical acquaintance with qualia.

Overcoming this ubiquitous temptation is the task of the next section, which will seek to establish the unsalvageable incoherence of the hunches that lead to the paradox by looking more closely at their sources and their motivation.

4. Making Mistakes about Qualia

The idea that people might be mistaken about their own qualia is at the heart of the ongoing confusion, and must be explored in more detail, and with somewhat more realistic examples, if we are to see the delicate role it plays.

Intuition pump #7: Chase and Sanborn. Once upon a time there were two coffee tasters, Mr Chase and Mr Sanborn, who worked for Maxwell House.[7] Along with half a dozen other coffee tasters, their job was to ensure that the taste of Maxwell House stayed constant, year after year. One day, about six years after Mr Chase had come to work for Maxwell House, he confessed to Mr Sanborn:

> I hate to admit it, but I'm not enjoying this work any more. When I came to Maxwell House six years ago, I thought Maxwell House coffee was the best-tasting coffee in the world. I was proud to have a share in the responsibility for preserving that flavor over the years. And we've done our job well; the coffee tastes just the same today as it tasted when I arrived. But, you know, I no longer like it! My tastes have changed. I've become a more sophisticated coffee drinker. I no longer like *that taste* at all.

Sanborn greeted this revelation with considerable interest. 'It's funny you should mention it,' he replied, 'for something rather similar has happened to me.' He went on:

> When I arrived here, shortly before you did, I, like you, thought Maxwell House coffee was tops in flavor. And now I, like you, really don't care for the coffee we're making. But *my* tastes haven't changed; my . . . *tasters* have changed. That is, I think something has gone wrong

with my taste buds or some other part of my taste-analyzing perceptual machinery. Maxwell House coffee doesn't taste to me the way it used to taste; if only it did, I'd still love it, for I still think *that taste* is the best taste in coffee. Now I'm not saying we haven't done our job well. You other tasters all agree that the taste is the same, and I must admit that on a day-to-day basis I can detect no change either. So it must be my problem alone. I guess I'm no longer cut out for this work.

Chase and Sanborn are alike in one way at least: they both used to like Maxwell House coffee, and now neither likes it. But they claim to be different in another way. Maxwell House tastes to Chase just the way it always did, but not so for Sanborn. But can we take their protestations at face value? Must we? Might one or both of them simply be wrong? Might their predicaments be importantly the same and their apparent disagreement more a difference in manner of expression than in experiential or psychological state? Since both of them make claims that depend on the reliability of their memories, is there any way to check on this reliability?

My reason for introducing two characters in the example is not to set up an interpersonal comparison between how the coffee tastes to Chase and how it tastes to Sanborn, but just to exhibit, side-by-side, two poles between which cases of intrapersonal experiential shift can wander. Such cases of intrapersonal experiential shift, and the possibility of adaptation to them, or interference with memory in them, have often been discussed in the literature on qualia, but without sufficient attention to the details, in my opinion. Let us look at Chase first. Falling in for the nonce with the received manner of speaking, it appears at first that there are the following possibilities:

(a) Chase's coffee-taste-qualia have stayed constant, while his reactive attitudes to those qualia, devolving on his canons of aesthetic judgment, etc., have shifted—which is what he seems, in his informal, casual way, to be asserting.
(b) Chase is simply wrong about the constancy of his qualia; they have shifted gradually and imperceptibly over the years, while his standards of taste haven't budged—in spite of his delusions about having become more sophisticated. He is in the state Sanborn claims to be in, but just lacks Sanborn's self-knowledge.
(c) Chase is in some predicament intermediate between (a) and (b); his qualia have shifted some *and* his standards of judgment have also slipped.

Sanborn's case seems amenable to three counterpart versions:

(a) Sanborn is right; his qualia have shifted, due to some sort of derangement in his perceptual machinery, but his standards have indeed remained constant.
(b) Sanborn's standards have shifted unbeknownst to him. He is thus misremembering his past experiences, in what we might call a nostalgia effect. Think of the familiar experience of returning to some object from your childhood (a classroom desk, a tree-house) and finding it much smaller than you remember it to have been. Presumably as you grew larger your internal standard for what was large grew with you somehow, but your memories (which are stored as fractions or multiples of that standard) didn't compensate, and hence when you consult your memory, it returns a distorted judgment. Sanborn's nostalgia-tinged memory of good old Maxwell House is similarly distorted. (There are obviously many different ways this impressionistic sketch of a memory mechanism could be implemented, and there is considerable experimental work in cognitive psychology that suggests how different hypotheses about such mechanisms could be tested.)
(c) As before, Sanborn's state is some combination of (a) and (b).

I think that everyone writing about qualia today would agree that there are all these possibilities for Chase and Sanborn. I know of no one these days who is tempted to defend the high line on infallibility or incorrigibility that would declare that alternative (a) is—and must be—the truth in each case, since people just cannot be wrong about such private, subjective matters.[8]

Since quandaries are about to arise, however, it might be wise to review in outline why the attractiveness of the infallibilist position is only superficial, so it won't recover its erstwhile allure when the going gets tough. First, in the wake of Wittgenstein 1958 and Malcolm 1956, 1959, we have seen that one way to buy such infallibility is to acquiesce in the complete evaporation of content (Dennett 1976). 'Imagine someone saying: 'But I know how tall

I am!' and laying his hand on top of his head to prove *it*' (Wittgenstein 1958, p. 96). By diminishing one's claim until there is nothing left to be right or wrong about, one can achieve a certain empty invincibility, but that will not do in this case. One of the things we want Chase to be right about (if he is right) is that he is not in Sanborn's predicament, so if the claim is to be viewed as infallible, it can hardly be because it declines to assert anything.

There is a strong temptation, I have found, to respond to my claims in this paper more or less as follows: 'But after all is said and done, there is still something I know in a special way: I know *how it is with me right now.*' But if absolutely nothing follows from this presumed knowledge—nothing, for instance, that would shed any light on the different psychological claims that might be true of Chase or Sanborn—what is the point of asserting that one has it? Perhaps people just want to reaffirm their sense of proprietorship over their own conscious states.

The infallibilist line on qualia treats them as properties of one's experience one cannot in principle misdiscover, and this is a mysterious doctrine (at least as mysterious as papal infallibility) unless we shift the emphasis a little and treat qualia as *logical constructs* out of subjects' qualia-judgments: a subject's experience has the quale *F* if and only if the subject judges his experience to have quale *F*. We can then treat such judgings as constitutive acts, in effect, bringing the quale into existence by the same sort of license as novelists have to determine the hair color of their characters by fiat. We do not ask how Dostoevski knows that Raskolnikov's hair is light brown.

There is a limited use for such interpretations of subjects' protocols, I have argued (Dennett 1978a; 1979, especially pp. 109–10; 1982), but they will not help the defenders of qualia here. Logical constructs out of judgments must be viewed as akin to theorists' fictions, and the friends of qualia want the existence of a particular quale in any particular case to be an empirical fact in good standing, not a theorist's useful interpretive fiction, else it will not loom as a challenge to functionalism or materialism or third-person, objective science.

It seems easy enough, then, to dream up empirical tests that would tend to confirm Chase and Sanborn's different tales, but if passing such tests could support their authority (that is to say, their reliability), failing the tests would have to undermine it. The price you pay for the possibility of empirically confirming your assertions is the outside chance of being discredited. The friends of qualia are prepared, today, to pay that price, but perhaps only because they haven't reckoned how the bargain they have struck will subvert the concept they want to defend.

Consider how we could shed light on the question of where the truth lies in the particular cases of Chase and Sanborn, even if we might not be able to settle the matter definitively. It is obvious that there might be telling objective support for one extreme version or another of their stories. Thus if Chase is unable to reidentify coffees, teas, and wines in blind tastings in which only minutes intervene between first and second sips, his claim to *know* that Maxwell House tastes just the same to him now as it did six years ago will be seriously undercut. Alternatively, if he does excellently in blind tastings, and exhibits considerable knowledge about the canons of coffee style (if such there be), his claim to have become a more sophisticated taster will be supported. Exploitation of the standard principles of inductive testing—basically Mill's method of differences—can go a long way toward indicating what sort of change has occurred in Chase or Sanborn—a change near the brute perceptual processing end of the spectrum or a change near the ultimate reactive judgment end of the spectrum. And as Shoemaker 1982 and others have noted, physiological measures, suitably interpreted in some larger theoretical framework, could also weight the scales in favor of one extreme or the other. For instance, the well-studied phenomenon of induced illusory boundaries (see figure 25.1) has often been claimed to be a particularly 'cognitive' illusion, dependent on 'top down' processes, and hence, presumably, near the reactive judgment end of the spectrum, but recent experimental work (Von der Heydt et al. 1984) has revealed that 'edge detector' neurons *relatively* low in the visual pathways—in area 18 of the visual cortex—are as responsive to illusory edges as to real light–dark boundaries on the retina, suggesting (but not quite proving, since these might somehow still be 'descending effects') that illusory contours are not imposed from on high, but generated quite early in visual processing. One can imagine discovering a similarly 'early' anomaly in the pathways leading from taste buds to judgment in Sanborn, for instance, tending to confirm his claim that he has suffered some change in his basic perceptual—as opposed to judgmental—machinery.

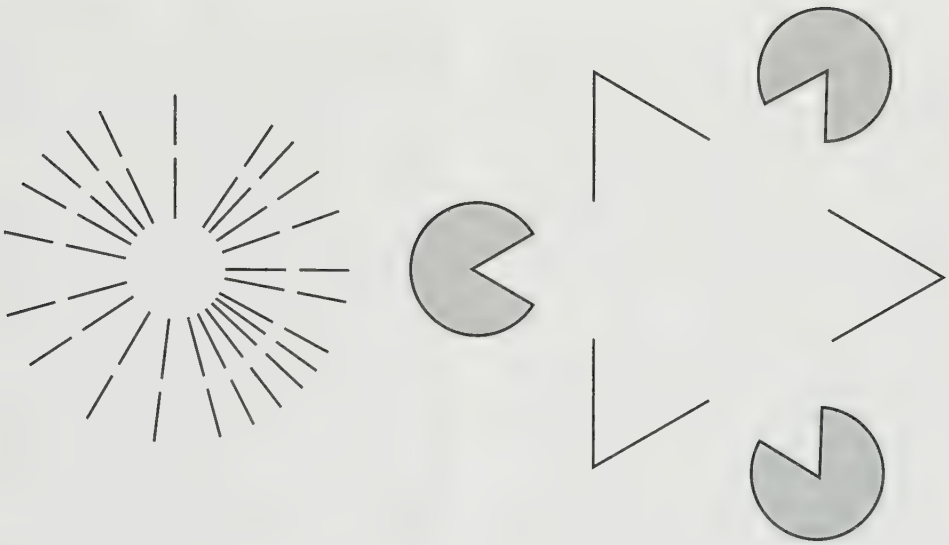


Figure 25.1

But let us not overestimate the resolving power of such empirical testing. The space in each case between the two poles represented by possibility (a) and possibility (b) would be occupied by phenomena that were the product, somehow, of two factors in varying proportion: roughly, dispositions to generate or produce qualia and dispositions to react to the qualia once they are produced. (That is how our intuitive picture of qualia would envisage it.) Qualia are supposed to affect our action or behavior only via the intermediary of our judgments about them, so any behavioral test, such as a discrimination or memory test, since it takes acts based on judgments as its primary data, can give us direct evidence only about the *resultant* of our two factors. In extreme cases we can have indirect evidence to suggest that one factor has varied a great deal, the other factor hardly at all, and we can test the hypothesis further by checking the relative sensitivity of the subject to variations in the conditions that presumably alter the two component factors. But such indirect testing cannot be expected to resolve the issue when the effects are relatively small—when, for instance, our rival hypotheses are Chase's preferred hypothesis (a) and the minor variant to the effect that his qualia have shifted *a little* and his standards *less than he thinks*. This will be true even when we include in our data any unintended or unconscious behavioral effects, for their import will be ambiguous (Would a longer response latency in Chase today be indicative of a process of 'attempted qualia renormalization' or 'extended aesthetic evaluation'?)

The limited evidential power of neurophysiology comes out particularly clearly if we imagine a case of adaptation. Suppose, in *intuition pump #8: the gradual post-operative recovery*, that we have somehow 'surgically inverted' Chase's taste bud connections in the standard imaginary way: post-operatively, sugar tastes salty, salt tastes sour, etc. But suppose further—and this is as realistic a supposition as its denial—that Chase has subsequently compensated—as revealed by his behavior. He now *says* that the sugary substance we place on his tongue is sweet, and no longer favors gravy on his ice cream. Let us suppose the compensation is so thorough that on all behavioral and verbal tests his performance is indistinguishable from that of normal subjects—and from his own presurgical performance.

If all the internal compensatory adjustment has been accomplished early in the process—intuitively, pre-qualia—then his qualia today are restored to just as they were (relative to external sources of stimulation) before the surgery. If on the other hand some or all of the internal compensatory adjustment is post-qualia, then his qualia have not been renormalized *even if he thinks they have*. But the physiological facts will not in themselves shed any light on where in the stream of physiological process twixt tasting and telling to draw the line at which the putative qualia appear as properties of that phase of the process. The qualia are the 'immediate or phenomenal' properties, of course, but this description will not serve to locate the right phase in the physiological stream, for, echoing

intuition pump #6, there will always be at least two possible ways of interpreting the neurophysiological theory, however it comes out. Suppose our physiological theory tells us (in as much detail as you like) that the compensatory effect in him has been achieved by an *adjustment in the memory-accessing process* that is required for our victim to compare today's hues to those of yore. There are *still* two stories that might be told:

I. Chase's current qualia are still abnormal, but thanks to the revision in his memory-accessing process, he has in effect adjusted his memories of how things used to taste, so he no longer notices any anomaly.

II. The memory-comparison step occurs just prior to the qualia phase in taste perception; thanks to the revision, it now *yields* the same old qualia for the same stimulation.

In (I) the qualia contribute to the input, in effect, to the memory-comparator. In (II) they are part of the output of the memory-comparator. These seem to be two substantially different hypotheses, but the physiological evidence, no matter how well developed, will not tell us on which side of memory to put the qualia. Chase's introspective evidence will not settle the issue between (I) and (II) either, since *ex hypothesi* those stories are not reliably distinguishable by him. Remember that it was in order to confirm or disconfirm Chase's opinion that we turned to the neurophysiological evidence in the first place. We can hardly use his opinion in the end to settle the matter between our rival neurophysiological theories. Chase may think that he thinks his experiences are the same as before *because* they really are (and he remembers accurately how it used to be), but he must admit that he has no introspective resources for distinguishing that possibility from alternative (I), on which he thinks things are as they used to be *because* his memory of how they used to be has been distorted by his new compensatory habits.

Faced with their subject's systematic neutrality, the physiologists may have their own reasons for preferring (I) to (II) or vice versa, for they may have *appropriated* the term 'qualia' to their own theoretical ends, to denote some family of detectable properties that strike them as playing an important role in their neurophysiological theory of perceptual recognition and memory. Chase or Sanborn might complain—in the company of more than a few philosophical spokesmen—that these properties the neurophysiologists choose to call 'qualia' are not the qualia they are speaking of. The scientists' retort is: 'If we cannot distinguish (I) from (II), we certainly cannot support either of your claims. If you want our support, you must relinquish your concept of qualia.'

What is striking about this is not just that the empirical methods would fall short of distinguishing what seem to be such different claims about qualia, but that they would fall short *in spite of being better evidence than the subject's own introspective convictions.* For the subject's own judgments, like the behaviors or actions that express them, are the resultant of our two postulated factors, and cannot discern the component proportions any better than external behavioral tests can. Indeed, a subject's 'introspective' convictions will generally be *worse* evidence than what outside observers can gather. For if our subject is—as most are—a 'naive subject,' unacquainted with statistical data about his own case or similar cases, his immediate, frank judgments are, evidentially, like any naive observer's perceptual judgments about factors in the outside world. Chase's intuitive judgments about his qualia constancy are no better off, epistemically, than his intuitive judgments about, say, lighting intensity constancy or room temperature constancy—or his own body temperature constancy. Moving to a condition inside his body does not change the intimacy of the epistemic relation in any special way. Is Chase running a fever or just feeling feverish? Unless he has taken steps to calibrate and cross-check his own performance, his opinion that his fever-perception apparatus is undisturbed is no better than a hunch. Similarly, Chase may have a strongly held opinion about the degree to which his taste-perceiving apparatus has maintained its integrity, and the degree to which his judgment has evolved through sophistication, but pending the results of the sort of laborious third-person testing just imagined, he would be a fool to claim to know—especially to know directly or immediately—that his was a pure case (a), closer to (a) than to (b), or a case near (b).

He is on quite firm ground, epistemically, when he reports that *the relation* between his coffee-sipping activity and his judging activity has changed. Recall that this is the factor that Chase and Sanborn have in common: they used to like Maxwell House; now they don't. But unless he carries out on himself the sorts of tests others might carry out on him, his convictions about what has stayed constant (or nearly so) and what has shifted *must be sheer guessing.*

But then qualia—supposing for the time being that we know what we are talking about—must lose one of their 'essential' second-order properties: far from being directly or immediately apprehensible properties of our experience, they are properties whose changes or constancies are either entirely beyond our ken, or inferrable (at best) from 'third-person' examinations of our behavioral and physiological reaction patterns (if Chase and Sanborn acquiesce in the neurophysiologists' sense of the term). On this view, Chase and Sanborn should be viewed not as introspectors capable of a privileged view of these properties, but as autopsychologists, theorists whose convictions about the properties of their own nervous systems are based not only on their 'immediate' or current experiential convictions, but also on their appreciation of the import of events they remember from the recent past.

There are, as we shall see, good reasons for neurophysiologists and other 'objective, third-person' theorists to single out such a class of properties to study. But they are not qualia, for the simple reason that one's epistemic relation to them is *exactly* the same as one's epistemic relation to such external, but readily—if fallibly—detectable, properties as room temperature or weight. The idea that one should consult an outside expert, and perform elaborate behavioral tests on oneself in order to confirm what qualia one had, surely takes us too far away from our original idea of qualia as properties with which we have a particularly intimate acquaintance.

So perhaps we have taken a wrong turning. The doctrine that led to this embarrassing result was the doctrine that sharply distinguished qualia from their (normal) effects on reactions. Consider Chase again. He claims that coffee tastes 'just the same' as it always did, but he admits—nay insists—that his reaction to 'that taste' is not what it used to be. That is, he pretends to be able to divorce his apprehension (or recollection) of the quale—the taste, in ordinary parlance—from his different reactions to the taste. But this apprehension or recollection is itself a reaction to the presumed quale, so some sleight-of-hand is being perpetrated—innocently no doubt—by Chase. So suppose instead that Chase had insisted that precisely *because* his reaction was now different, the taste had changed for him. (When he told his wife his original tale, she said 'Don't be silly! Once you add the dislike you change the experience!'—and the more he thought about it, the more he decided she was right.)

Intuition pump #9: the experienced beer drinker. It is familiarly said that beer, for example, is an acquired taste; one gradually trains oneself—or just comes—to enjoy that flavor. What flavor? The flavor of the first sip? No one could like *that* flavor, an experienced beer drinker might retort:

> Beer tastes different to the experienced beer drinker. If beer went on tasting to me the way the first sip tasted, I would never have gone on drinking beer! Or to put the same point the other way around, if my first sip of beer had tasted to me the way my most recent sip just tasted, I would never have had to acquire the taste in the first place! I would have loved the first sip as much as the one I just enjoyed.

If we let this speech pass, we must admit that beer is *not* an acquired taste. No one comes to enjoy *the way the first sip tasted.* Instead, prolonged beer drinking leads people to experience a taste they enjoy, but precisely their enjoying the taste guarantees that it is not the taste they first experienced.[9]

But this conclusion, if it is accepted, wreaks havoc of a different sort with the traditional philosophical view of qualia. For if it is admitted that one's attitudes towards, or reactions to, experiences are in any way and in any degree constitutive of their experiential qualities, so that a change in reactivity *amounts to* or *guarantees* a change in the property, then those properties, those 'qualitative or phenomenal features,' cease to be 'intrinsic' properties, and in fact become paradigmatically extrinsic, relational properties.

Properties that 'seem intrinsic' at first often turn out on more careful analysis to be relational. Bennett 1965 is the author of *intuition pump #10: the world-wide eugenics experiment.* He draws our attention to phenol-thio-urea, a substance which tastes very bitter to three-fourths of humanity, and as tasteless as water to the rest. Is it bitter? Since the reactivity to phenol-thio-urea is genetically transmitted, we could make it paradigmatically bitter by performing a large-scale breeding experiment: prevent the people to whom it is tasteless from breeding, and in a few generations phenol would be as bitter as anything to be found in the world. But we could also (in principle!) perform the contrary feat of mass 'eugenics' and thereby make phenol paradigmatically tasteless—as tasteless as water—without ever touching phenol. Clearly, public bitterness or tastelessness is not an intrinsic property of phenol-thio-urea but a relational property, since the property is

changed by a change in the reference class of normal detectors.

The public versions of perceptual 'qualia' all *seem* intrinsic, in spite of their relationality. They are not alone. Think of the 'felt value' of a dollar (or whatever your native currency is). 'How much is that in *real* money?' the American tourist is reputed to have asked, hoping to translate a foreign price onto the scale of 'intrinsic value' he keeps in his head. As Elster 1985 claims, 'there is a tendency to overlook the implicitly relational character of certain monadic predicates.' Walzer 1985 points out that '. . . a ten-dollar bill might seem to have a life of its own as a thing of value, but, as Elster suggests, its value implicitly depends on 'other people who are prepared to accept money as payment for goods.' But even as one concedes this, there is still a tendency to reserve something subjective, felt value, as an 'intrinsic' property of that ten-dollar bill. But as we now see, such intrinsic properties cannot be properties to which a subject's access is in any way privileged.

Which way should Chase go? Should he take his wife's advice and declare that since he can't stand the coffee any more, it no longer tastes the same to him (it used to taste good and now it tastes bad)? Or should he say that really, in a certain sense, it does taste the way it always did or at least it sort of does—when you subtract the fact that it tastes so bad now, of course?

We have now reached the heart of my case. The fact is that we have to ask Chase which way he wants to go, and there really are two drastically different alternatives available to him *if we force the issue*. Which way would *you* go? Which concept of qualia did you 'always have in the back of your mind,' guiding your imagination as you thought about theories? If you acknowledge that the answer is not obvious, and especially if you complain that this forced choice drives apart two aspects that you had supposed united in your pretheoretic concept, you support my contention that there is no secure foundation in ordinary 'folk psychology' for a concept of qualia. We *normally* think in a confused and potentially incoherent way when we think about the ways things seem to us.

When Chase thinks of 'that taste' he thinks equivocally or vaguely. He harkens back in memory to earlier experiences but need not try—or be able—to settle whether he is including any or all of his reactions or excluding them from what he intends by 'that taste.' His state then and his state now are different—*that* he can avow with confidence—but he has no 'immediate' resources for making a finer distinction, nor any need to do so.[10]

This suggests that qualia are no more essential to the professional vocabulary of the phenomenologist (or professional coffee taster) than to the vocabulary of the physiologist (Dennett 1978b). To see this, consider again the example of my dislike of cauliflower. Imagine now, in *intuition pump #11: the cauliflower cure*, that someone offers me a pill to cure my loathing for cauliflower. He promises that after I swallow this pill cauliflower will taste exactly the same to me as it always has, but I will like that taste! 'Hang on,' I might reply. 'I think you may have just contradicted yourself.' But in any event I take the pill and it works. I become an instant cauliflower-appreciater, but if I am asked which of the two possible effects (Chase-type or Sanborn-type) the pill has had on me, I will be puzzled, and will find nothing *in my experience* to shed light on the question. Of course I recognize that the taste is (sort of) the same—the pill hasn't made cauliflower taste like chocolate cake, after all—but at the same time my experience is so different now that I resist saying that cauliflower tastes the way it used to taste. There is in any event no reason to be cowed into supposing that my cauliflower experiences have some intrinsic properties behind, or in addition to, their various dispositional, reaction-provoking properties.

'But in principle there has to be a right answer to the question of how it is, intrinsically, with you now, even if you are unable to say with any confidence!' Why? Would one say the same about all other properties of experience? Consider *intuition pump #12: visual field inversion created by wearing inverting spectacles*, a phenomenon which has been empirically studied for years. (G. M. Stratton published the pioneering work in 1896, and J. J. Gibson and Ivo Kohler were among the principal investigators. For an introductory account, see Gregory 1977.) After wearing inverting spectacles for several days subjects make an astonishingly successful adaptation. Suppose we pressed on them this question: 'Does your adaptation consist in your re-inverting your visual field, or in your turning the rest of your mind upside-down in a host of compensations?' If they demur, may we insist that there has to be a right answer, even if they cannot say with any confidence which it is? Such an insistence would lead directly to a new version of the old inverted spectrum thought experiment: 'How do I know whether some people see things upside-down (but are

perfectly used to it), while others see things right-side-up?'

Only a very naive view of visual perception could sustain the idea that one's visual field has a property of right-side-upness or upsidedownness *independent of one's dispositions to react to it*—'intrinsic right-side-upness' we could call it. (See my discussion of the properties of the 'images' processed by the robot SHAKEY in Dennett 1982.) So not all properties of conscious experience invite or require treatment as 'intrinsic' properties. Is there something distinguishing about a certain subclass of properties (the 'qualitative or phenomenal' subclass, presumably) that forces us to treat them—unlike subjective right-side-upness—as intrinsic properties? If not, such properties have no role to play, in either physiological theories of experience, or in introspective theories.

Some may be inclined to argue this way: I can definitely imagine the experience of 'spectrum inversion' from the inside; after all I have actually experienced temporary effects of the same type, such as the 'taste displacement' effect of the maple syrup on the orange juice. What is imaginable, or actual, is possible. Therefore spectrum inversion or displacement (in all sensory modalities) is possible. But such phenomena just *are* the inversion or displacement of qualia, or intrinsic subjective properties. Therefore there must be qualia: intrinsic subjective properties.

This is fallacious. What one imagines and what one says one imagines may be two different things. To imagine visual field inversion, of the sort Stratton and Kohler's subjects experienced, is not necessarily to imagine the absolute inversion of a visual field (even if that is what it 'feels like' to the subjects). Less obviously, imagining—as vividly as you like—a case of subjective color-perception displacement is not necessarily imagining what that phenomenon is typically called by philosophers: an inverted or displaced spectrum *of qualia.* In so far as that term carries the problematic implications scouted here, there is no support for its use arising simply from the vividness or naturalness of the imagined possibility.

If there are no such properties as qualia, does that mean that 'spectrum inversion' is impossible? Yes and no. Spectrum inversion as classically debated is impossible, but something like it is perfectly possible—something that is as like 'qualia inversion' as visual field inversion is like the impossible *absolute* visual image inversion we just dismissed.

5. Some Puzzling Real Cases

It is not enough to withhold our theoretical allegiances until the sunny day when the philosophers complete the tricky task of purifying the everyday concept of qualia. Unless we take active steps to shed this source concept, and replace it with better ideas, it will continue to cripple our imaginations and systematically distort our attempts to understand the phenomena already encountered.

What we find, if we look at the actual phenomena of anomalies of color perception, for instance, amply bears out our suspicions about the inadequacy of the traditional notion of qualia. Several varieties of *cerebral achromatopsia* (brain based impairment of color vision) have been reported, and while there remains much that is unsettled about their analysis, there is little doubt that the philosophical thought experiments have underestimated or overlooked the possibilities for counter-intuitive collections of symptoms, as a few very brief excerpts from case histories will reveal.

> Objects to the right of the vertical meridian appeared to be of normal hue, while to the left they were perceived only in shades of gray, though without distortions of form. . . . He was unable to recognize or name any color in any portion of the left field of either eye, including bright reds, blues, greens and yellows. As soon as any portion of the colored object crossed the vertical meridian, he was able to instantly recognize and accurately name its color. (Damasio et al. 1980)

This patient would seem at first to be unproblematically describable as suffering a shift or loss of color qualia in the left hemifield, but there is a problem of interpretation here, brought about by another case:

> The patient failed in all tasks in which he was required to match the seen color with its spoken name. Thus, the patient failed to give the names of colors and failed to choose a color in response to its name. By contrast he succeeded on all tasks where the matching was either purely verbal or purely nonverbal. Thus, he could give verbally the names of colors corresponding to named objects and vice versa. He could match seen colors to each other and to pictures of objects and could sort colors without error. (Geschwind and Fusillo 1966)

This second patient was quite unaware of any deficit. He 'never replied with a simple 'I don't know' to the demand for naming a color' (Geschwind and Fusillo 1966, p. 140). There is a striking contrast between these two patients;

both have impaired ability to name the colors of things in at least part of their visual field, but whereas the former is acutely aware of his deficit, the latter is not. Does this difference make all the difference about qualia? If so, what on earth should we say about this third patient?

> His other main complaint was that 'everything looked black or grey' and this caused him some difficulty in everyday life. . . . He had considerable difficulty recognizing and naming colours. He would, for example, usually describe bright red objects as either red or black, bright green objects as either green, blue or black, and bright blue objects as black. The difficulty appeared to be perceptual and he would make remarks suggesting this; for example when shown a bright red object he said 'a dirty smudgy red, not as red as you would normally see red.' Colours of lesser saturation or brightness were described in such terms as 'grey' 'off-white' or 'black,' but if told to guess at the colour, he would be correct on about 50 per cent of occasions, being notably less successful with blues and greens than reds. (Meadows 1974)

This man's awareness of his deficit is problematic to say the least. It contrasts rather sharply with yet another case:

> One morning in November 1977, upon awakening, she noted that although she was able to see details of objects and people, colors appeared 'drained out' and 'not true.' She had no other complaint . . . her vision was good, 20/20 in each eye . . . The difficulty in color perception persisted, and she had to seek the advice of her husband to choose what to wear. Eight weeks later she noted that she could no longer recognize the faces of her husband and daughter. . . . [So in] addition to achromatopsia, the patient had prosopagnosia, but her linguistic and cognitive performances were otherwise unaffected. The patient was able to tell her story cogently and to have remarkable insight about her defects. (Damasio et al. 1980)

As Meadows notes, 'Some patients thus complain that their vision for colours is defective while others have no spontaneous complaint but show striking abnormalities on testing.'

What should one say in these cases? When no complaint is volunteered but the patient shows an impairment in color vision, is this a sign that his qualia are unaffected? ('His capacities to discriminate are terribly impaired, but luckily for him, his inner life is untouched by this merely public loss!') We could line up the qualia this way, but equally we could claim that the patient has simply not noticed the perhaps gradual draining away or inversion or merging of his qualia revealed by his poor performance. ('So slowly did his inner life lose its complexity and variety that he never noticed how impoverished it had become!') What if our last patient described her complaint just as she did above, but performed normally on testing? One hypothesis would be that her qualia had indeed, as she suggested, become washed out. Another would be that in the light of her sterling performance on the color discrimination tests, her qualia were fine; she was suffering from some hysterical or depressive anomaly, a sort of color-vision hypochondria that makes her complain about a loss of color perception. Or perhaps one could claim that her qualia were untouched; her disorder was purely verbal: an anomalous understanding of the words she uses to describe her experience. (Other startlingly specific color-*word* disorders have been reported in the literature.)

The traditional concept leads us to overlook genuine possibilities. Once we have learned of the curious deficit reported by Geschwind and Fusillo, for instance, we realize that our first patient was never tested to see if he could still sort colors seen on the left or pass other non-naming, non-verbal color-blindness tests. Those tests are by no means superfluous. Perhaps he would have passed them; perhaps, *in spite of what he says* his qualia are as intact for the left field as for the right!—if we take the capacity to pass such tests as 'criterial.' Perhaps his problem is 'purely verbal.' If your reaction to this hypothesis is that this is impossible, that must mean you are making his verbal, reporting behavior sovereign in settling the issue—but then you must rule out a priori the possibility of the condition I described as color-vision hypochondria.

There is no prospect of *finding* the answers to these brain-teasers in our everyday usage or the intuitions it arouses, but it is of course open to the philosopher to *create* an edifice of theory defending a particular set of interlocking proposals. The problem is that although normally a certain family of stimulus and bodily conditions yields a certain family of effects, any particular effect can be disconnected, and our intuitions do not tell us which effects are 'essential' to quale identity or qualia constancy (cf. Dennett 1978a, chapter 11). It seems fairly obvious to me that none of the real problems of interpretation that face us in these curious cases is advanced by any analysis of how the concept of *qualia* is to be applied—unless we wish to propose a novel, technical sense for which the traditional term might be appropriated. But that

would be at least a tactical error: the intuitions that surround and *purport* to anchor the current understanding of the term are revealed to be in utter disarray when confronted with these cases.

My informal sampling shows that some philosophers have strong opinions about each case and how it should be described in terms of qualia, but they find they are in strident (and ultimately comic) disagreement with other philosophers about how these 'obvious' descriptions should go. Other philosophers discover they really don't know what to say—not because there aren't enough facts presented in the descriptions of the cases, but because it begins to dawn on them that they haven't really known what they were talking about over the years.

6. Filling the Vacuum

If qualia are such a bad idea, why have they seemed to be such a good idea? Why does it seem as if there are these intrinsic, ineffable, private, 'qualitative' properties in our experience? A review of the presumptive second-order properties of the properties of our conscious experiences will permit us to diagnose their attractiveness and find suitable substitutes. (For a similar exercise see Kitcher 1979).

Consider 'intrinsic' first. It is far from clear what an intrinsic property would be. Although the term has had a certain vogue in philosophy, and often seems to secure an important contrast, there has never been an accepted definition of the second-order property of intrinsicality. If even such a brilliant theory-monger as David Lewis can try and fail, by his own admission, to define the extrinsic/intrinsic distinction coherently, we can begin to wonder if the concept deserves our further attention after all. In fact Lewis 1983 begins his survey of versions of the distinction by listing as one option: 'We could Quine the lot, give over the entire family as unintelligible and dispensable,' but he dismisses the suggestion immediately: 'That would be absurd' (p. 197). In the end, however, his effort to salvage the accounts of Chisholm 1976 and Kim 1982 are stymied, and he conjectures that 'if we still want to break in we had best try another window' (p. 200).

Even if we are as loath as Lewis is to abandon the distinction, shouldn't we be suspicious of the following curious fact? If challenged to explain the idea of an intrinsic property to a neophyte, many people would hit on the following sort of example: consider Tom's ball; it has many properties, such as its being made of rubber from India, its belonging to Tom, its having spent the last week in the closet, and its redness. All but the last of these are clearly *relational* or *extrinsic* properties of the ball. Its redness, however, is an intrinsic property. Except this isn't so. Ever since Boyle and Locke we have known better. Redness—public redness—is a quintessentially relational property, as many thought experiments about 'secondary qualities' show. (One of the first was Berkeley's [1713] pail of lukewarm water, and one of the best is Bennett's [1965] phenol-thio-urea.) The seductive step, on learning that public redness (like public bitterness, etc.) is a relational property after all, is to cling to intrinsicality ('*something* has to be intrinsic!') and move it into the subject's head. It is often thought, in fact, that if we take a Lockean, relational position on objective bitterness, redness, etc., we *must* complete our account of the relations in question by appeal to non-relational, intrinsic properties. If what it is to be objectively bitter is to produce a certain effect in the members of the class of normal observers, we must be able to specify that effect, and distinguish it from the effect produced by objective sourness and so forth.

What else could distinguish this effect but some intrinsic property? Why not another relational or extrinsic property? The relational treatment of monetary value does not require, for its completion, the supposition of items of intrinsic value (value independent of the valuers' dispositions to react behaviorally). The claim that certain perceptual properties are different is, in the absence of any supporting argument, just question-begging. It will not do to say that it is just obvious that they are intrinsic. It may have seemed obvious to some, but the considerations raised by Chase's quandary show that it is far from obvious that any intrinsic property (whatever that comes to) could play the role of anchor for the Lockean relational treatment of the public perceptual properties.

Why not give up intrinsicality as a second-order property altogether, at least pending resolution of the disarray of philosophical opinion about what intrinsicality might be? Until such time the insistence that qualia are the intrinsic properties of experience is an empty gesture at best; no one could claim that it provides a clear, coherent, understood prerequisite for theory.[11]

What, then, of ineffability? Why does it seem that our conscious experiences have ineffable properties? Because they do have *practically* ineffable properties. Suppose, in *intuition*

pump #13: the osprey cry, that I have never heard the cry of an osprey, even in a recording, but know roughly, from reading my bird books, what to listen for: 'a series of short, sharp, cheeping whistles, *cheep cheep* or *chewk chewk*, etc; sounds annoyed' (Peterson 1947) (or words to that effect or better). The verbal description gives me a partial confinement of the logical space of possible bird cries. On its basis I can rule out many bird calls I have heard or might hear, but there is still a broad range of discriminable-by-me possibilities within which the actuality lies hidden from me like a needle in a haystack.

Then one day, armed with both my verbal description and my binoculars, I identify an osprey visually, and then hear its cry. So *that's* what it sounds like, I say to myself, ostending—it seems—a particular mental complex of intrinsic, ineffable qualia. I dub the complex 'S' (*pace* Wittgenstein), rehearse it in short term memory, check it against the bird book descriptions, and see that while the verbal descriptions are true, accurate and even poetically evocative—I decide I could not do better with a thousand words—they still fall short of *capturing* the qualia-complex I have called *S*. In fact, that is why I need the neologism '*S*' to refer directly to the ineffable property I cannot pick out by description. My perceptual experience has pinpointed for me the location of the osprey cry in the logical space of possibilities in a way verbal description could not.

But tempting as this view of matters is, it is overstated. First of all, it is obvious that from a single experience of this sort I don't—can't—know how to generalize to other osprey calls. Would a cry that differed only in being half an octave higher also be an osprey call? That is an empirical, ornithological question for which my experience provides scant evidence. But moreover—and this is a psychological, not ornithological, matter—I don't and can't know, from a single such experience, which physical variations and constancies in stimuli would produce an indistinguishable experience in me. Nor can I know whether I would react the same (have the same experience) if I were presented with what was, by all physical measures, a re-stimulation identical to the first. I cannot know the modulating effect, if any, of variations in my body (or psyche).

This inscrutability of projection is surely one of the sources of plausibility for Wittgenstein's skepticism regarding the possibility of a private language.

> Wittgenstein emphasizes that ostensive definitions are always in principle capable of being misunderstood, even the ostensive definition of a color word such as 'sepia.' How someone understands the word is exhibited in the way someone goes on, 'the use that he makes of the word defined.' One may go on in the right way given a purely minimal explanation, while on the other hand one may go on in another way no matter how many clarifications are added, since these too can be misunderstood . . . (Kripke 1982, p. 83; see also pp. 40–6)

But what is inscrutable in a single glance, and somewhat ambiguous after limited testing, can come to be justifiably seen as the deliverance of a highly specific, reliable, and projectible property-detector, once it has been field-tested under a suitably wide variety of circumstances.

In other words, when first I hear the osprey cry, I may have identified a property-detector in myself, but I have no idea (yet) what property my new-found property-detector detects. It might seem then that I know nothing new at all—that my novel experience has not improved my epistemic predicament in the slightest. But of course this is not so. I may not be able to describe the property or identify it relative to any readily usable public landmarks (yet), but I am acquainted with it in a modest way: I can refer to the property I detected: it is the property I detected in *that* event. My experience of the osprey cry has given me a new way of thinking about osprey cries (an unavoidably inflated way of saying something very simple) which is practically ineffable both because it has (as yet for me) an untested profile in response to perceptual circumstances, and because it is—as the poverty of the bird-book description attests—such a highly informative way of thinking: a deliverance of an informationally very sensitive portion of my nervous system.

In this instance I mean information in the formal information theory sense of the term. Consider (*intuition pump #14: the Jell-O box*) the old spy trick, most famously encountered in the case of Julius and Ethel Rosenberg, of improving on a password system by tearing something in two (a Jell-O box, in the Rosenberg's case), and giving half to each of the two parties who must be careful about identifying each other. Why does it work? Because tearing the paper in two produces an edge of such informational complexity that it would be virtually impossible to reproduce by deliberate construction. (Cutting the Jell-O box with straight edge and razor would entirely defeat the purpose.)

The particular jagged edge of one piece becomes a *practically* unique pattern-recognition device for its mate; it is an apparatus for detecting the shape property *M*, where *M* is uniquely instantiated by its mate. It is of the essence of the trick that we cannot replace our dummy predicate '*M*' with a longer, more complex, but accurate and exhaustive description of the property, for if we could, we could use the description as a recipe or feasible algorithm for producing another instance of *M* or another *M* detector. The only *readily available* way of saying what property *M* is is just to point to our *M*-detector and say that *M* is the shape property detected by this thing here.

And that is just what we do when we seem to ostend, with the mental finger of inter intention, a quale or qualia-complex in our experience. We refer to a property—a public property of uncharted boundaries—via reference to our personal and idiosyncratic capacity to respond to it. That idiosyncrasy is the extent of our privacy. If I wonder whether your blue is my blue, your middle-C is my middle-C, I can coherently be wondering whether our discrimination profiles over a wide variation in conditions will be approximately the same. And they may not be; people experience the world quite differently. But that is empirically discoverable by all the usual objective testing procedures.[12]

Peter Bieri has pointed out to me that there is a natural way of exploiting Dretske's (1981) sense of information in a reformulation of my first three second-order properties of qualia: intrinsicality, ineffability, and privacy. (There are problems with Dretske's attempt to harness information theory in this way—see my discussion in 'Evolution, error and intentionality' (Dennett 1987)—but they are not relevant to this point.) We could speak of what Bieri would call 'phenomenal information properties' of psychological events. Consider the information—what Dretske would call the *natural meaning*—that a type of internal perceptual event might carry. That it carries that information is an objective (and hence, in a loose sense, intrinsic) matter since it is independent of what information (if any) the subject *takes* the event type to carry. Exactly what information is carried is (practically) ineffable, for the reasons just given. And it is private in the sense just given: proprietary and potentially idiosyncratic.

Consider how Bieri's proposed 'phenomenal information properties' (let's call them *pips*) would apply in the case of Chase and Sanborn. Both Chase and Sanborn ought to wonder whether their pips have changed. Chase's speech shows that he is under the impression that his pips are unchanged (under normal circumstances—all bets are off if he has just eaten horseradish). He believes that the same objective things in the world—in particular, chemically identical caffeine-rich fluids—give rise to his particular types of taste-experiences now as six years ago.

Sanborn is under the impression that his pips are different. He thinks his objective property-detectors are deranged. He no longer has confidence that their deliverances today inform him of what they did six years ago. And what, exactly, did they inform him of then? If Sanborn were an ordinary person, we would not expect him to have an explicit answer, since most of us treat our taste-detectors as mere *M*-detectors, detecting whatever-it-is that they detect. (There are good reasons for this, analyzed by Akins 1987.) But professional coffee-tasters are probably different. They probably have some pretty good idea of what kind of chemical-analysis transduction machinery they have in their mouths and nervous systems.

So far, so good. We could reinterpret Chase and Sanborn's speeches as hypotheses about the constancies or changes in the outputs of their perceptual information-processing apparatus, and just the sort of empirical testing we imagined before would tend to confirm or disconfirm their opinions thus interpreted. But what would justify calling such an information-bearing property 'phenomenal'?

Such a pip has, as the testimony of Chase and Sanborn reveals, the power to provoke in Chase and Sanborn acts of (apparent) re-identification or recognition. This power is of course a Lockean, dispositional property on a par with the power of bitter things to provoke a certain reaction in people. It is this power alone, however it might be realized in the brain, that gives Chase and Sanborn 'access' to the deliverances of their individual property-detectors.

We may 'point inwardly' to one of the deliverances of our idiosyncratic, proprietary property-detectors, but when we do, what are we pointing *at?* What does that deliverance itself *consist of?* Or what are its consciously apprehensible properties, if not just our banished friends the qualia? We must be careful here, for if we invoke an inner perceptual process in which we observe the deliverance with some inner eye and thereby discern its properties, we will be stepping back into the frying pan of the view according to which qualia are just ordinary properties of our inner states.

But nothing requires us to make such an invocation. We don't have to know how we identify or re-identify or gain access to such internal response types in order to be able so to identify them. This is a point that was forcefully made by the pioneer functionalists and materialists, and has never been rebutted (Farrell 1950; Smart 1959). The properties of the 'thing experienced' are not to be confused with the properties of the event that realizes the experiencing. To put the matter vividly, the physical difference between someone's imagining a purple cow and imagining a green cow *might* be nothing more than the presence or absence of a particular zero or one in one of the brain's 'registers.' Such a brute physical presence is all that it would take to anchor the sorts of dispositional differences between imagining a purple cow and imagining a green cow that could then flow, causally, from that 'intrinsic' fact. (I doubt that this is what the friends of qualia have had in mind when they have insisted that qualia are intrinsic properties.)

Moreover, it is our very inability to expand on, or modify, these brute dispositions so to identify or recognize such states that creates the doctrinal illusion of 'homogeneity' or 'atomicity to analysis' or 'grainlessness' that characterizes the qualia of philosophical tradition.

This putative grainlessness, I hypothesize, is nothing but a sort of functional invariability: it is close kin to what Pylyshyn 1980, 1984 calls *cognitive impenetrability.* Moreover, this functional invariability or impenetrability is not absolute but itself plastic over time. Just as on the efferent side of the nervous system, *basic actions*—in the sense of Danto 1963, 1965 and others (see Goldman 1970)—have been discovered to be variable, and subject under training to decomposition (one can learn with the help of 'biofeedback' to will the firing of a particular motor neuron 'directly'), so what counts for an individual as the simple or atomic properties of experienced items is subject to variation with training.[13]

Consider the results of 'educating' the palate of a wine taster, or 'ear training' for musicians. What had been 'atomic' or 'unanalyzable' becomes noticeably compound and describable; pairs that had been indistinguishable become distinguishable, and when this happens we say *the experience changes.* A swift and striking example of this is illustrated in *intuition pump #15: the guitar string.* Pluck the bass or low E string open, and listen carefully to the sound. Does it have describable parts or is it one and whole and ineffably guitarish? Many will opt for the latter way of talking. Now pluck the open string again and carefully bring a finger down lightly over the octave fret to create a high 'harmonic.' Suddenly a *new* sound is heard: 'purer' somehow and of course an octave higher. Some people insist that this is an entirely novel sound, while others will describe the experience by saying 'the bottom fell out of the note'—leaving just the top. But then on a third open plucking one can hear, with surprising distinctness, the harmonic overtone that was isolated in the second plucking. The homogeneity and ineffability of the first experience is gone, replaced by a duality as 'directly apprehensible' and clearly describable as that of any chord.

The difference in experience is striking, but the complexity apprehended on the third plucking was *there* all along (being responded to or discriminated). After all, it was by the complex pattern of overtones that you were able to recognize the sound as that of a guitar rather than a lute or harpsichord. In other words, although the subjective experience has changed dramatically, the *pip* hasn't changed; you are still responding, as before, to a complex property so highly informative that it practically defies verbal description.

There is nothing to stop further refinement of one's capacity to describe this heretofore ineffable complexity. At any time, of course, there is one's current horizon of distinguishability—and that horizon is what sets, if anything does, what we should call the primary or atomic properties of what one consciously experiences (Farrell 1950). But it would be a mistake to transform the fact that inevitably there is a limit to our capacity to describe things we experience into the supposition that there are absolutely indescribable properties in our experience.

So when we look one last time at our original characterization of qualia, as ineffable, intrinsic, private, directly apprehensible properties of experience, we find that there is nothing to fill the bill. In their place are relatively or practically ineffable public properties we can refer to indirectly via reference to our private property-detectors—private only in the sense of idiosyncratic. And in so far as we wish to cling to our subjective authority about the occurrence within us of states of certain types or with certain properties, we can have some authority—not infallibility or incorrigibility, but something better than sheer guessing—but only if we restrict ourselves to relational, extrinsic properties like the power of certain internal states of ours to provoke acts of apparent re-identification. So contrary to what seems obvious at first blush, there simply are no qualia at all.[14]

REFERENCES

Akins, K., *Information and Organisms: Or, Why Nature Doesn't Build Epistemic Engines*, PhD dissertation, University of Michigan, Department of Philosophy, 1987.

Armstrong, D. and Malcolm, N., eds., *Consciousness and Causality* (Oxford: Basil Blackwell, 1984).

Bennett, J., "Substance, reality and primary qualities," *American Philosophical Quarterly* 2 (1965): pp. 1–17.

Berkeley, G. *Three Dialogues between Hylas and Philonous* (1713).

Block, N., "Troubles with Functionalism," in *Perception and Cognition: Minnesota Studies in the Philosophy of Science*, Vol. IX, W. Savage, ed., (Minneapolis: University of Minnesota Press, 1978).

_____ "Are absent qualia impossible?" *Philosophical Review* 89 (1980): p. 257.

Block, N., and Fodor, J., "What psychological states are not," *Philosophical Review* 81 (1972): pp. 159–81.

Chisholm, R. *Person and Object* (La Salle, IL: Open Court Press, 1976).

Churchland, P. M. *Scientific Realism and the Plasticity of Mind* (Cambridge, UK: Cambridge University Press, 1979).

Churchland, P. M., "Reduction, qualia and the direct inspection of brain states," *Journal of Philosophy* LXXXII (1985): pp. 8–28.

Conee, E., "The possibility of absent qualia," *Philosophical Review* 94 (1985): pp. 345–66.

Damasio, A., et al., "Central Achromatopsia: Behavioral, anatomic, and physiological aspects," *Neurology* 30 (1980): pp. 1064–71.

Danto, A., "What we can do," *Journal of Philosophy*, LX (1963): pp. 435–45.

_____ "Basic actions," *American Philosophical Quarterly* (1965): pp. 141–48.

Davis, L., "Functionalism and absent qualia," *Philosophical Studies* 41 (1982): pp. 231–51.

Dennett, D. C. *Content and Consciousness* (London: Routledge & Kegan Paul, 1969).

_____ "Are dreams experiences?" *Philosophical Review* 85, (1976): pp. 151–71. (Reprinted in Dennett 1978a.)

_____ *Brainstorms, Philosophical Essays on Mind and Psychology* (Cambridge, MA: Bradford Books/ MIT Press, 1978a).

_____ "Two approaches to mental images," in Dennett, *Brainstorms*.

_____*The Philosophical Lexicon* (1978c), 8th ed.

_____ "On the absence of phenomenology," in *Body, Mind, and Method (Festschrift for Virgil Aldrich)*, D. F. Gustafson and B. L. Tapscott, eds., (Dordrecht: Reidel, 1979): pp. 93–114.

_____ "Wondering where the yellow went," *Monist* 64 (1981): pp. 102–8.

_____ "How to study human consciousness empirically: Or nothing comes to mind," *Synthese* 53 (1982): pp. 159–80.

_____*The Intentional Stance* (Cambridge, MA: Bradford Books, MIT Press, 1987).

Dretske, F. *Knowledge and the Flow of Information* (Cambridge, MA: Bradford Books, MIT Press, 1981).

Elster, J. *Making Sense of Marx* (Cambridge, UK: Cambridge University Press, 1985).

Farrell, B. A., "Experience," *Mind* 59 (1950): PP. 170–98.

Gert, B., "Imagination and verifiability," *Philosophical Studies* 16 (1965): pp. 44–47.

Geschwind, N., and Fusillo, M., "Color-naming defects in association with alexia," *Archives of neurology* 15 (1966): pp. 137–46.

Goldman, A. *A Theory of Human Action* (Englewood Cliffs, NJ: Prentice Hall, 1970).

Gregory, R. *Eye and Brain* (London: Weidenfeld & Nicolson, 1977), 3rd ed.

Hofstadter, D., and Dennett, D. C. *The Mind's I: Fantasies and Reflections on Mind and Soul* (New York: Basic Books, 1981).

Kim, J., "Psychophysical supervenience," *Philosophical Studies* 41 (1982): pp. 51–70.

Kitcher, P., "Phenomenal qualities," *American Philosophical Quarterly* 16 (1979): pp. 123–29.

Kripke, S. *Wittgenstein on Rules and Private Language* (Cambridge, MA: Harvard University Press, 1982).

Lehrer, A. *Wine and Conversation* (Bloomington: University of Indiana Press, 1983).

Lewis, D. "Extrinsic properties," *Philosophical Studies* 44 (1983): pp. 197–200.

Locke, J. *An Essay Concerning Human Understanding*, A. C. Fraser, ed., (New York: Dover, 1959).

Lycan, W., "Inverted spectrum," *Ratio* XV (1973): pp. 315–19.

Malcolm, N., "Dreaming and skepticism," *Philosophical Review* 64 (1956): pp. 14–37.

_____ *Dreaming* (London: Routledge & Kegan Paul, 1959).

Meadows, J. C., "Disturbed perception of colours associated with localized cerebral lesions," *Brain* 97 (1974): pp. 615–32.

Millikan, R. *Language, Thought and Other Biological Categories* (Cambridge, MA: Bradford Books, MIT Press, 1984).

Nagel, T. "What is it like to be a bat?" *Philosophical Review* 83 (1974): pp. 435–51.

_____ *The View from Nowhere* (Oxford: Oxford University Press, 1986).

Peirce, C. *Collected Works*, C. Hartshorne and P. Weiss, eds. (Cambridge, MA: Harvard University Press, 1931–58.

Peterson, R. T. *A Field Guide to the Birds* (Boston: Houghton Mifflin, 1947).

Putnam, H., "Brains and behavior," in *Analytical Philosophy*, J. Butler, ed., (Oxford: Basil Blackwell, 1965), second series.

Pylyshyn, Z., "Computation and cognition: Issues in the foundation of cognitive science," *Behavioral and Brain Sciences* 3 (1980): pp. 111–32.

_____ *Computation and Cognition: Toward a Foundation for Cognitive Science* (Cambridge, MA: Bradford Books, MIT Press, 1984).

Sellars, W., "Foundations for a metaphysics of pure process" (the Carus Lectures), *Monist* 64 (1981): pp. 3–90.

Shoemaker, S., "Time without change," *Journal of Philosophy* 66 (1969): pp. 363–81.

_____ "Functionalism and qualia," *Philosophical Studies* 27 (1975): pp. 291–315.

_____ "Absent qualia are impossible—A Reply to Block," *Philosophical Review* 90 (1981): pp. 581–99.

_____ "The inverted spectrum," *Journal of Philosophy* 79 (1982): pp. 357–81.

_____"Postscript (1983)," in *Identity, Cause, and Mind* (Cambridge, UK: Cambridge University Press, 1984), pp. 351–57.

Smart, J. J. C., "Sensations and brain processes," *Philosophical Review* 68 (1959): pp. 141–56. (Reprinted in Chappell 1962).

Stich, S. *From Folk Psychology to Cognitive Science: The Case against Belief* (Cambridge, MA: Bradford Books, MIT Press, 1983).

Taylor, D. M., "The incommunicability of content," *Mind* 75 (1966): pp. 527–41.

Von der Heydt, R., Peterhans, E., and Baumgartner, G., "Illusory contours and cortical neuron response," *Science* 224 (1984): pp. 1260–262.

Walzer, M., "What's left of Marx," *New York Review of Books,* Nov. 21, 1985, pp. 43–46.

White, S., "Professor Shoemaker and so-called 'qualia' of experience," *Philosophical Studies* 47 (1985): pp. 369–83.

Wittgenstein, L. *Philosophical Investigations,* G. E. M. Anscombe, ed., (Oxford: Basil Blackwell, 1958).

NOTES

1. A representative sample of the most recent literature on qualia would include Block 1980; Shoemaker 1981, 1982; Davis 1982; White 1985; Armstrong and Malcolm 1984; Churchland 1985; and Conee 1985.
2. The difference between 'eliminative materialism'—of which my position on qualia is an instance—and a 'reductive' materialism that takes on the burden of identifying the problematic item in terms of the foundational materialistic theory is thus often best seen not so much as a doctrinal issue as a tactical issue: how might we most gracefully or effectively enlighten the confused in this instance? See my discussion of 'fatigues' in the Introduction to *Brainstorms* (Dennett 1978a), and earlier, my discussion of what the enlightened ought to say about the metaphysical status of *sakes* and *voices* in *Content and Consciousness* (Dennett 1969), chapter 1.
3. The plausibility of this concession depends less on a high regard for the technology than on a proper skepticism about human powers, now documented in a fascinating study by Lehrer 1983.
4. Shoemaker 1984 seems to be moving reluctantly towards agreement with this conclusion: 'So unless we can find some grounds on which we can deny the possibility of the sort of situation envisaged . . . we must apparently choose between rejecting the functionalist account of qualitative similarity and rejecting the standard conception of qualia.I would prefer not to have to make this choice; but if I am forced to make it, I reject the standard conception of qualia' (p. 356).
5. Shoemaker 1982 attributes a view to Wittgenstein (acknowledging that 'it is none too clear' that this is actually what Wittgenstein held) which is very close to the view I defend here. But to Shoemaker, 'it would seem offhand that Wittgenstein was mistaken' (p. 360), a claim Shoemaker supports with a far from offhand thought experiment—which Shoemaker misanalyzes if the present paper is correct. (There is no good reason, contrary to Shoemaker's declaration, to believe that his subject's *experience* is systematically different from what it was before the inversion.) Smart 1959 expresses guarded and partial approval of Wittgenstein's hard line, but cannot see his way clear to as uncompromising an eliminativism as I maintain here.
6. In 1979, I read an earlier version of this paper in Oxford, with a commentary by John Foster, who defended qualia to the last breath, which was: 'qualia should not be quined but fostered!' Symmetry demands, of course, the following definition for the eighth edition of *The Philosophical Lexicon:* 'foster, *v.* To acclaim resolutely the existence or importance of something chimerical or insignificant.'
7. This example first appeared in print in my 'Reflections on Smullyan' in *The Mind's I* (Hofstadter and Dennett 1981), pp. 427–28.
8. Kripke 1982 comes close, when he asks rhetorically 'Do I not know, directly, and *with a fair degree of certainty* [emphasis added], that I mean plus [by the function I call 'plus']?' (p. 40). Kripke does not tell us what is implied by 'a fair degree of certainty,' but presumably he means by this remark to declare his allegiance to what Millikan 1984 attacks under the name of 'meaning rationalism.'
9. We can save the traditional claim by ignoring presumably private or subjective qualia and talking always of public tastes—such as the public taste of Maxwell House coffee that both Chase and Sanborn agree has remained constant. Individuals can be said to acquire a taste for such a public taste.
10. 'I am not so wild as to deny that my sensation of red today is like my sensation of red yesterday. I only say that the similarity can *consist* only in the physiological force behind consciousness—which leads me to say, I recognize this feeling the same as the former one, and so does not consist in a community of sensation.' (C. S. Peirce, *Collected Works,* vol. V, p. 172, fn. 2).
11. A heroic (and, to me, baffling) refusal to abandon intrinsicality is Wilfrid Sellars's contemplation over the years of his famous pink ice cube, which leads him to postulate a revolution in microphysics, restoring objective 'absolute sensory processes' in the face of Boyle and Locke and almost everybody since them. See Sellars (1981) and my commentary (Dennett 1981).

12. Stich 1983 discusses the implications for psychological theory of incommensurability problems that can arise from such differences in discrimination profiles. See especially chapters 4 and 5.
13. See Churchland 1979, especially chapter 2, for supporting observations on the variability of perceptual properties, and for novel arguments against the use of 'intrinsic properties' as determiners of the meaning of perceptual predicates. See also Churchland 1985 for further arguments and observations in support of the position sketched here.
14. The first version of this paper was presented at University College London, in November 1978, and in various revisions at a dozen other universities in 1979 and 1980. It was never published, but was circulated widely as Tufts University Cognitive Science Working Paper #7, December 1979. A second version was presented at the Universities of Adelaide and Sydney in 1984, and in 1985 to psychology department colloquia at Harvard and Brown under the title 'Properties of conscious experience.' The second version was the basis for my presentation at the workshop on consciousness in modern science, Villa Olmo, Como, Italy, April 1985, and circulated in preprint in 1985, again under the title 'Quining qualia.' The present version, the fourth, is a substantial revision, thanks to the helpful comments of many people, including Kathleen Akins, Ned Block, Alan Cowey, Sydney Shoemaker, Peter Bieri, William Lycan, Paul Churchland, Gilbert Harman, and the participants at Villa Olmo.

Explaining Consciousness

David M. Rosenthal

Among mental phenomena, none seems so thoroughly to resist informative explanation as does consciousness. Part of the difficulty is due to our using the term 'conscious' and its cognates to cover several distinct phenomena, whose connections with one another are not always clear. And that often leads us to run these distinct phenomena together. Any attempt to explain consciousness, therefore, must begin by distinguishing the various things we call consciousness.

One such phenomenon is closely related to simply being awake. We describe people, and other creatures, as being conscious when they are awake and their sensory systems are receptive in the way normal for a waking state. I call this phenomenon *creature consciousness.* Consciousness in this sense is a biological matter, consisting in a creature's not being unconscious—that is, roughly, in its not being asleep or knocked out.

But we also use the term 'consciousness' for other phenomena that seem a lot less tractable to understanding and explanation. Not only do we distinguish between conscious and unconscious creatures; we also distinguish between mental states that are conscious and those which are not. I'll call this second property *state consciousness.* It's widely recognized that not all mental states are conscious. Intentional states such as beliefs and desires plainly occur without being conscious.[1] And, despite some division of opinion on the matter, I shall argue that the same is true of sensory states, such as pains and sensations of color. Such states not only can, but often do occur nonconsciously.[2]

Though creature consciousness and state consciousness are distinct properties, they are very likely related in various ways. Perhaps, for example, creatures must themselves be conscious for any of their mental states to be conscious, though if ordinary dreams are ever conscious states they are counterexamples to this generalization.[3]

Whatever the case about that, the property of creature consciousness is relatively unproblematic. We can see this by considering creatures mentally less well-endowed than we are whose mental states are never conscious, even when they are awake.[4] Their mental states are all like the nonconscious mental states we are in when we are awake. Doubtless some creatures are actually like this, say, frogs or turtles. And it's plain that when none of a creature's mental states is conscious, there is nothing puzzling about what it is for the creature to be conscious. Some theorists might deny that such a case is possible, urging that no creature counts

First published in the first edition of this book: D. J. Chalmers, ed., *Philosophy of Mind: Classical and Contemporary Readings* (Oxford University Press, 2002), pp. 406–21.

as conscious unless some of the mental states it is in are conscious states. But this seems little more than an unwarranted extrapolation of the normal human waking state to the case of all creatures. Even if their view were correct, moreover, it would be state consciousness that introduces the apparent mystery.

What is puzzling about consciousness must therefore be a matter not of creature consciousness, but of the consciousness of a creature's mental states. Because creature consciousness involves being responsive to sensory stimuli, if sensory states were all conscious, every conscious creature would perforce be in some conscious states. But it would still, then, be the consciousness of the states, not of the creature, which seems to induce some mystery.

For this reason, I shall focus here on state consciousness. After laying some groundwork in section I, I go on in section II to develop a hypothesis about what it is for a mental state to be conscious. On this hypothesis, a mental state is conscious if it is accompanied by a specific type of thought. This is so whether the state that is conscious is itself an intentional state or a sensory state. Section III, then, supports this hypothesis with an argument that appeals to the ability creatures like ourselves have to report noninferentially about their own conscious states.

Sections IV and V take up the special case of conscious qualitative states. I argue in section IV that such sensory consciousness is just a special case of state consciousness and poses no additional problems of its own. And section V gives reasons for thinking that an accompanying intentional state can actually result in there being something it's like for one to be in a conscious sensory state. Section VI, finally, considers two general questions about state consciousness: What function it might have and whether consciousness can misrepresent what mental states we are in.

I. State Consciousness and Transitive Consciousness

Whatever else we may discover about consciousness, it's clear that, if one is totally unaware of some mental state, that state is not a conscious state. A state may of course be conscious without one's paying conscious attention to it and, indeed, even without one's being conscious of every mental aspect of the state. But if one is not at all aware of a state, that state is not a conscious state. This observation provides a useful start toward a theory of state consciousness. Because it is sufficient for a state not to be conscious that one be completely unaware of it, being aware of a state is perforce a necessary condition for that state to be a conscious state.

Being aware of a mental state, however, is not also a sufficient condition for the state to be conscious. There are ways we can be aware of our mental states even when those states are not conscious states. So, if we can rule out those ways, we'll be left with the particular way in which we are aware of our mental states when those states are conscious states. And this would give us a condition that's both necessary and sufficient for a mental state to be conscious.

For present purposes, I'll speak interchangeably of being aware of something and being conscious of that thing. So my strategy is to explain a state's being a conscious state in terms of our being conscious of that state in some particular way. No circle is involved here, since we are explaining one phenomenon in terms of another. It is one thing for us to be conscious *of* something—what we may call *transitive consciousness*—and another for a state to be a conscious state—what I'm calling state consciousness. And we understand transitive consciousness—our being conscious *of* things—independently of understanding what it is for mental states to be conscious states. We are transitively conscious of something by virtue of being either in an intentional or a sensory state whose content is directed upon that thing. And a state's having a certain content is a distinct property from that of a state's being conscious.[5]

It seems relatively uncontroversial that a state of which one is in no way transitively conscious could not be a conscious state. Even Descartes' usage, which still strongly influences our own, conforms to this commonsense observation, since he invariably describes the states we call conscious as states we are immediately conscious of. Nonetheless, Fred Dretske has recently challenged the observation that we are conscious of all our conscious states. According to Dretske, a state's being conscious does not consist in one's being conscious of the state; rather, a state is conscious if, in virtue of being in that state, one is conscious of something or conscious that something is the case. But every mental state satisfies this condition; so Dretske must hold that all mental states are conscious states. Accordingly, he urges that alleged cases of nonconscious mental states are

unconvincing. Thus it is often said that a long-distance driver whose attention lapses perceives the road unconsciously,[6] but Dretske rightly notes that perceiving can be inattentive without failing to be conscious.[7]

Many other examples of nonconscious mental states, however, are far more decisive. We often consciously puzzle over a question about what to do or how to solve a problem, only to have the answer occur to us later, without the matter having in the meantime been in any way consciously before our mind. Though it doesn't seem, from a first-person point of view, that we were thinking about the issue, it's clear that we must have been. And unlike the case of the long-distance driver, here no shift of attention would change things. Also we often take in sensory information without being at all aware of doing so, again no matter what we're paying attention to. Since, from a first-person perspective, we seem not to be in any relevant sensory states, those states are not conscious states.

Dretske also argues, however, that there are actual counterexamples to the idea that we are transitively conscious of all our conscious states. To adapt his argument slightly, consider two scenes, one of ten trees and the other just like it, but with one tree missing. And suppose that I consciously see both scenes, and indeed that I consciously see all the trees in each scene. But suppose, finally, that despite all this I do not notice any difference between the two scenes.

Dretske sensibly assumes that in this case I have conscious experiences of both scenes, including all the trees in each. Moreover, there is some part of the conscious experience of ten trees that is not part of the conscious experience of nine trees. That part is itself a conscious experience—a conscious experience of a tree. But, because I am not transitively conscious of the difference between the two scenes, Dretske concludes that I will not be transitively conscious of the experience of that extra tree. If so, the experience of the extra tree is a conscious experience of which I am not transitively conscious.[8]

This sort of thing is hardly an esoteric occurrence. Indeed, it happens all the time; let one scene be a slightly later version of the other, such that the later scene is altered in some small, unnoticed way. So, if Dretske's argument is sound, we often fail to be conscious of our conscious experiences.[9]

But the argument isn't sound. One can be conscious of an experience in one respect while not being conscious of it in another. For example, one may be conscious of a visual experience as an experience of a blurry patch, but not as an experience of a particular kind of object. Similarly, one could be conscious of the experience of the extra tree as an experience of a tree, or even just as part of one's overall experience, without being at all conscious of it as the thing that makes the difference between the experiences of the two scenes. Presumably this is just what happens in the case Dretske constructs. Dretske has not described a conscious state of which one is not transitively conscious.

There is a complication in Dretske's discussion that is worth noting. Dretske insists that being conscious of a difference, unlike being conscious of concrete objects and events, always amounts to being conscious 'that such a difference exists.'[10] So he might urge that being conscious of a difference is always being conscious of it as a difference. But this won't help. The experience of the extra tree is that in virtue of which the two overall experiences differ. Still, one can be conscious of the thing in virtue of which they happen to differ without being conscious that they do differ. As Dretske would put it, one can be conscious of that in virtue of which they differ but not of the difference between them;[11] indeed, he explicitly acknowledges that this very thing can happen.[12] Dretske's argument does not, therefore, undermine the commonsense observation that we are transitively conscious of all our conscious states.[13]

II. The Hypothesis

Let us turn, then, to the question of what it is that is special about the way we are transitively conscious of our mental states when those states are conscious states. Perhaps the most obvious thing is that, when a state is conscious, we are conscious of it in a way that seems immediate. Descartes emphasized this intuitive immediacy,[14] which many have thought points toward a Cartesian theory of mind, on which a mental state's being conscious is an intrinsic property of that state. If nothing mediates between a state and one's being transitively conscious of it, perhaps that transitive consciousness is something internal to the state itself.

But the intuition about immediacy does not show that a mental state's being conscious is internal to the state. It does seem, from a first-person point of view, that nothing mediates

between the conscious states we are conscious of and our transitive consciousness of them. But all that shows is that, if anything does mediate between a conscious state and our transitive consciousness of it, the mediating factor is not one we are conscious of. And the absence of conscious mediation is no reason to think that nonconscious mediation does not occur.[15] Failure to appreciate this has led some to hold that we are conscious of our conscious states in a way wholly unlike the way we are conscious of everything else.

Even when something mediates between a conscious mental state and our being conscious of it, we can be conscious of the mediating factor; we just cannot be conscious of it as mediating. Compare what happens in perceiving. When we consciously perceive things, our conscious sensory states mediate between our perceptions and the objects we perceive, and since those states are conscious, we are conscious of them. Still, nothing in these cases seems intuitively to mediate. That's because we aren't conscious of anything as mediating. And the best explanation of that, in turn, is that the conscious sensory states that do in fact mediate do not figure in any conscious inference on which our perceiving is based. Similarly with the way we are conscious of our conscious mental states. Our being conscious of them seems unmediated because we are conscious of them in a way that relies on no conscious inference, no inference, that is, of which we are aware.[16]

Consider a case. I am annoyed, but unaware of it. Though my annoyance is not conscious, you observe my annoyed behavior and tell me I am annoyed. There are two ways I might react. I might accept what you tell me, but still feel no conscious annoyance. My belief that I'm annoyed would be the result of a conscious inference based on your remark, and possibly also a conscious inference from my coming to notice my own relevant behavior.[17] But there is another possibility; your remark might cause me to become conscious of my annoyance independently of any such conscious inference. In that case my annoyance would have become a conscious state.

A state's being conscious involves one's being noninferentially conscious of that state. Can we pin down any further the way we are transitively conscious of our conscious states? There are two broad ways of being transitively conscious of things. We are conscious of something when we see it or hear it, or perceive it in some other way. And we are conscious of something when we have a thought about it. Which kind of transitive consciousness is relevant here? When our mental states are conscious, do we somehow sense those states or do we have thoughts about them?

The perceptual model may seem inviting. When we perceive things, we seem intuitively to be directly conscious of them; nothing seems to mediate between our perceptions and the objects we perceive.[18] So perhaps the perceptual model can explain the apparent immediacy of the way we are conscious of our conscious states. But this advantage of the perceptual model won't help us decide between that model and the alternative view that we are conscious of our conscious states by having thoughts about them. Even though our thoughts do often rely on conscious inferences involving perceptions or other thoughts, they often don't.

There is, however, another consideration that seems to favor the perceptual model. A theory of consciousness must explain the qualitative dimension of our conscious sensory states. And sensing always involves some sensory quality. So if we are conscious of our conscious states by sensing them, perhaps we can explain the qualitative dimension of consciousness as due to that higher-order sensing. Such an explanation, however, would at best just put off the problem, since the qualitative aspect of this higher-order perceiving would itself need to be explained in turn.

Not only do the considerations favoring the perceptual model fail to hold up; there is also reason to reject the model. Higher-order sensing would have to exhibit characteristic mental qualities; what qualities might those be? One possibility is that the higher-order perception and the state we perceive would both exhibit the same sensory quality. But this is theoretically unmotivated. When we perceive something, the quality of our perceptual state is distinct from any property of the object we perceive. When we see a tomato, for example, the redness of our sensation is not the same property as the redness of the tomato.[19] So we have no reason to think that the higher-order qualities would be the same as those of our lower-order states.

If the higher- and lower-order qualities were distinct, however, it's a mystery what those higher-order qualities could be. What mental qualities are there in our mental lives other than those which characterize our first-order sensory states? And if the higher-order qualities are neither the same as nor distinct from our first-order qualities, the higher-order states in virtue of

which we are conscious of our conscious states cannot have qualities at all. But if those higher-order states have no qualitative properties, they can only be higher-order intentional states of some sort.[20]

We must therefore reject the perceptual model of how we are transitively conscious of our conscious states. The only alternative is that we are conscious of our conscious states by virtue of having thoughts about them. Since these thoughts are about other mental states, I shall refer to them as *higher-order thoughts* (HOTs).

This narrows down somewhat the way we are transitively conscious of our mental states when those states are conscious. But we can narrow things down even more. When a mental state is conscious, we are conscious of being in that state; so the content of our HOT must be, roughly, that one is in that very state.[21] And, since merely being disposed to have a thought about something does not make one conscious of that thing, the HOT must be an occurrent thought, rather than just a disposition to think that one is in the target state. Moreover, when we are conscious of something by being in an intentional state that's about that thing, the intentional state is normally assertoric. Indeed, it's likely that being in an intentional state whose mental attitude is not assertoric does not result in one's being conscious of the thing the intentional state is about.[22] So we should require that the HOT has an assertoric mental attitude.[23] Finally, to capture the intuition about immediacy, we have seen that our HOTs must be independent of any inference of which we are aware. Our hypothesis, therefore, is that a mental state is conscious just in case it is accompanied by a noninferential, nondispositional, assertoric thought to the effect that one is in that very state.[24]

One problem that seems to face this hypothesis is that, even when we are in many conscious states, we are typically unaware of having any such HOTs. But this is not a difficulty; we are conscious of our HOTs only when those thoughts are themselves conscious, and it's rare that they are. Moreover, the hypothesis readily explains why this should be so. The HOTs it posits are conscious thoughts only when they are accompanied, in turn, by yet higher-order thoughts about them, and that seldom happens. Not having conscious HOTs, moreover, does nothing at all to show that we do not have HOTs that fail to be conscious.

There is another reason it's useful to distinguish cases in which HOTs are conscious from cases in which they are not. The way we are ordinarily conscious of our conscious states differs from the way we are conscious of mental states of which we are introspectively conscious. Being introspectively conscious of a mental state involves, roughly, our deliberately focusing on that state, and very few of our conscious states are the subjects of any such introspective scrutiny. If being conscious of a mental state were the same as being introspectively conscious of it, it would be rare that we are conscious of our conscious states, and we would be unable to explain state consciousness in terms of transitive consciousness. Not distinguishing the two, moreover, would lead one mistakenly to see the HOT hypothesis as providing a theory only of introspective consciousness, and not of state consciousness generally.[25] But the present hypothesis actually allows us to explain what is distinctive about introspective consciousness. A state is introspectively conscious when the accompanying HOT is a conscious thought. Ordinary, nonintrospective state consciousness, by contrast, occurs instead when the HOT is not itself conscious.

The HOT model is a hypothesis about the nature of state consciousness, not an analysis of that concept. So it doesn't count against the hypothesis simply that one can imagine its not holding; one can always imagine things being different from the way they are.

There is an especially interesting argument that supports the appeal to HOTs. When a mental state is conscious, one can noninferentially report being in that state, whereas one cannot report one's nonconscious mental states. Every speech act, moreover, expresses an intentional state with the same content as that of the speech act and a mental attitude that corresponds to its illocutionary force. So a noninferential report that one is in a mental state will express a non-inferential thought that one is in that state, that is, a HOT about the state. We can best explain this ability noninferentially to report our conscious states by supposing that the relevant HOT is there to be expressed. Correspondingly, the best explanation of our inability to report nonconscious states is that no HOTs accompany them.[26]

One might reply that the ability to report conscious states shows only that there is a disposition for these states to be accompanied by HOTs, not that any HOTs actually accompany them.[27] Indeed, Peter Carruthers has extensively developed and supported the view that conscious states are simply those disposed to

be accompanied by HOTs, and no actual HOT need occur. This, he argues, avoids having to posit the overwhelming computational capacity and cognitive space required for actual HOTs.[28]

But this concern is not compelling. Neural implementation is not a problem, since ample cortical resources exist to accommodate actual HOTs. And, though introspection seems to suggest that the mind cannot accommodate very many actual HOTs at a time, that worry is also groundless. Introspection can tell us only about our conscious states, and by hypothesis HOTs are seldom conscious.

In any case, the dispositional model cannot explain what it is for states to be conscious. A mental state's being conscious consists in one's being conscious of that state in some suitable way, and simply being disposed to have a thought about something cannot make one conscious of it. Carruthers urges that we can get around this difficulty if we understand a state's intentional content in terms of what other intentional states it is disposed to cause. A state's being disposed to cause a HOT might then confer suitable higher-order content on that state itself. But, if a state's being disposed to cause a HOT were a function of its intentional content, we could no longer explain how a state with some particular content is sometimes conscious and sometimes not.

III. Sensory Consciousness

On this argument, sensory consciousness is simply a special case of state consciousness—the special case in which the state that's conscious is a sensory state. Sensory states are states with sensory quality. So sensory consciousness occurs when a mental state has two properties: sensory quality and the property of state consciousness.

Moreover, these two properties are distinct and can occur independently of one another. State consciousness can of course occur without sensory quality, since nonsensory, intentional states are often conscious. But the converse is possible as well; sensory qualities can occur without state consciousness. Sensory qualities are just whatever properties sensory states have on the basis of which we distinguish among them and sort them into types. Since state consciousness consists in our being conscious of a mental state in some suitable way, these properties are independent of state consciousness. We would need some special reason to think that the properties on the basis of which we distinguish among sensations cannot occur except when we're conscious of the states that have those properties. It's hard to see what special reason there could be.

This conclusion conflicts with the familiar contention that sensory quality cannot occur nonconsciously. On that view, state consciousness is intrinsic, or essential, to sensory quality. But it's far from clear that this view is correct. Subliminal perception and peripheral vision both involve perceptual sensations of which we're wholly unaware, and the same is very likely true of such dissociative phenomena as blind-sight.[29] Bodily sensations such as pains can also occur without being conscious. For example, we often have a headache or other pain throughout an extended period even when distractions intermittently make us wholly unaware of the pain.

One could of course simply dig in one's heels and insist that these phenomena are mere physiological occurrences that instantiate no sensory quality, and therefore that they are not mental phenomena at all. But without independent argument, that move amounts simply to saving a view by verbal fiat.

In any case there is good reason to resist that claim. The relevant nonconscious phenomena occur as essential parts of distinctively mental processes, and that suggests that they are themselves mental phenomena.[30] More specifically, conscious sensory states play the same roles in mental processing when their sensory qualities are the same, and correspondingly different roles when the qualities differ. And the nonconscious states in subliminal perception, peripheral vision, and blindsight play roles that in some respects at least parallel the roles played by conscious sensory states.

When bodily and perceptual sensations occur consciously, we taxonomize them by way of the sensory qualities we are conscious of. What is it, then, in virtue of which we taxonomize the nonconscious states that occur in these cases? Since many of the same qualitative distinctions figure in the nonconscious cases as figure in conscious sensing, we must assume that the nonconscious cases have the very same qualitative properties.[31] Sensory qualities are the distinguishing properties of sensory states, the properties in virtue of which we classify those states. We use the properties we are conscious of to taxonomize sensory states generally, whether they are conscious or not. It's just that in the nonconscious cases we are not conscious

of those properties. And, since there is nothing problematic about these distinguishing properties when the states that have them are not conscious, there can be no reason to find those properties puzzling when we are conscious of them. Sensory qualities will seem mysterious only if we assume that they cannot occur without being conscious. These considerations make the claim that sensory quality must be conscious seem less like a compelling commonsense intuition than a question-begging theoretical doctrine.

There is, of course, nothing it's like to have a pain or a sensation of red unless the sensation in question is conscious. And some have argued from this to the conclusion that sensory quality simply cannot exist unless there's something it's like to have it.[32] But what it's like for one to have a pain, in the relevant sense of that idiom, is simply what it's like for one to be conscious of having that pain. So there won't be anything it's like to have a pain unless the pain is conscious. Of course, if nonconscious pains were impossible, there would be no difference between a pain's existing and its being conscious, and its sensory quality would then exist only when there is something it's like to have it. But it begs the question simply to assume that pains, or other sensations, cannot exist nonconsciously. Moreover, the intuition that sensory states cannot exist nonconsciously gets whatever force it has from our first-person point of view. And it's unreasonable to rely on consciousness to tell us whether some phenomenon can exist outside of consciousness.

In a useful series of papers, Ned Block has urged that there are two distinct properties of mental states, both of which we call consciousness. One is captured by the notion of there being something it's like for one to be in a particular mental state; Block calls this property *phenomenal consciousness.* A state has the other property when its content is 'poised to be used as a premise in reasoning, . . . [and] for [the] *rational* control of action and . . . speech.'[33] This second property Block calls *access consciousness.* And he maintains that the two properties are, conceptually at least, independent. If Block is right, there is no single property of state consciousness, and the kind of consciousness that is characteristic of sensory states is, conceptually at least, distinct from the kind exhibited by many nonsensory states.

The idea behind Block's account of access consciousness is that a state's playing various executive, inferential, and reporting roles involves one's having access to that state, and having access to a state makes it conscious. But that's not always the case. States often play executive, inferential, and even certain reporting roles[34] without being conscious in any intuitive sense whatever. So, for a state to be access conscious, one must have access to that state, presumably by being transitively conscious of it in an intuitively immediate way.[35]

Block's appeal to states' playing these roles doubtless reflects a desire to account for this kind of consciousness in computationally inspired functional terms, by providing a kind of flow chart that charts the connections a state has with various relevant systems. But for any such attempt to succeed, it must reflect an initial account of such consciousness in ordinary folk-psychological terms. Going straight to a subpersonal account is unlikely to give even an extensionally adequate account.

Block is doubtless right that access consciousness often occurs without phenomenal consciousness. We frequently have access to our mental states in the relevant way without there being anything it's like for us to be in them. Indeed, that's typically how it is with our thoughts and other intentional states. But the converse is far less clear. A state is access conscious only if one is transitively conscious of it. And if one is in no way transitively conscious of a mental state, there is nothing it's like for one to be in that state. It's not enough for the state just to have the distinguishing properties characteristic of some type of sensory state; for there to be something it's like for one to be in a state, one must be conscious of those distinguishing properties. So phenomenal consciousness cannot occur without access consciousness. Block's distinction does not, after all, show that sensory states are conscious in a way distinct from other types of mental state, nor that sensory states are in some special way invariably conscious.[36]

IV. HOTs and What It's Like

Nonetheless, there does seem to be a serious problem about what it is for sensory states to be conscious. When a sensory state is conscious, there is something it's like for us to be in that state. When it's not conscious, we do not consciously experience any of its qualitative properties; so then there is nothing it's like for us to be in that state. How can we explain this difference? A sensory state's being conscious means

that we are transitively conscious of that state in some suitable way. So being transitively conscious of a sensory state, in that particular way, must result in there being something it's like to be in that state. But how can being transitively conscious of a sensory state have this result? What way of being transitively conscious of our sensory states could, by itself, give rise to there being something it's like for us to be in those states? Perhaps, after all, Block is right that a sensory state's being conscious is not a matter of one's having suitable access to it.

The difficulty seems particularly pressing for the HOT hypothesis. An attraction of the perceptual model was that it might help explain the qualitative dimension of our conscious sensory states. Since perceiving involves sensory qualities, if a state's being conscious consisted in our perceiving it, perhaps we could explain the way we are conscious of the qualities of our conscious sensations. As we saw, that explanation fails, since the higher-order qualities it appeals to would themselves need to be explained. But the HOT hypothesis may seem even less well-suited to deal with this problem. How can one's being in an intentional state, of whatever sort, result in there being something it's like for one to be in a conscious sensory state?

There are two ways the HOT theorist might try to show that being in a suitable intentional state can have this result. One would be to show that it's evident, from a first-person point of view, that one has a suitable HOT when, and only when, there is something it's like for one to be in some sensory state. We could then argue that one's having that HOT is responsible for there being something it's like for one to be in that state.

But if the HOT hypothesis is correct, we cannot expect to find any such first-person correlations. That's because, on that hypothesis, the HOTs in virtue of which our sensory states are conscious are seldom conscious thoughts. And when a thought is not conscious, it will seem, from a first-person point of view, that one does not have it.

So if the HOT hypothesis is correct, it will rarely seem, from a first-person point of view, that HOTs accompany one's conscious sensory states. Our first-person access reveals correlations only with conscious HOTs, not HOTs generally. And HOTs are conscious only in those rare cases in which one has a third-order thought about the HOT. But on the HOT hypothesis, HOTs need not be conscious for there to be something it's like to be in the target sensory states. So we cannot hope to test the hypothesis by correlating in a first-person way the occurrence of HOTs with there being something it's like to be in conscious sensory states.

But we need not rely solely on first-person considerations; there are other factors that help establish the correlation between having HOTs and there being something it's like for one to be in conscious sensory states. In particular, there is a striking connection between what HOTs we are able to have and what sensory qualities we are able to be aware of. And the best explanation of this connection is that accompanying HOTs do result in there being something it's like for one to be in states with those sensory qualities.

Consider wine tasting. Learning new concepts for our experiences of the gustatory and olfactory properties of wines typically leads to our being conscious of more fine-grained differences among the qualities of our sensory states. Similarly with other sensory modalities; acquiring new concepts for specific musical and artistic experiences, for example, enables us to have conscious experiences with more finely differentiated sensory qualities. Somehow, the new concepts appear to generate new conscious sensory qualities.

There are two ways this might happen. One is that coming to have new concepts results in our sensory states' coming to have distinguishing properties that they did not previously have. This is highly implausible. How could merely having new concepts give rise to our sensory states' having new properties? On a widespread view, concepts are abilities to think certain things; how could having a new ability change the properties of the sensory states that result from the same type of stimulus?

But there is another possibility. The new concepts might result in new conscious qualities not by generating those properties, but by making us conscious of properties that were already there. The new concepts would enable us to be conscious of sensory qualities we already had, but had not been conscious of.[37]

Possessing a concept allows us to form intentional states that have a certain range of contents. So which contents our intentional states can have must somehow make a difference to which sensory qualities can occur consciously. Moreover, the new concepts, which make possible conscious experiences with qualities that seem new to us, are the concepts of those very qualities.[38] So being able to form intentional states about certain sensory qualities must

somehow result in our being able to experience those qualities consciously. It must result, that is, in there being something specific that it's like for us to be in the relevant sensory states.

How could this happen? The only plausible explanation is that a sensory quality's being conscious does actually consist in our having a HOT about that quality. This is true not only of the relatively finely differentiated qualities we have just now been considering. We can extrapolate to any sensory quality, however crudely individuated, and extrapolate even to whether or not we are conscious of any quality at all.

Take the conscious experience of hearing the sound of an oboe. If one's HOTs couldn't classify one's sensations in terms of the sound of an oboe but only that of some undifferentiated woodwind, having that sensation could not be for one like hearing an oboe. And if one also lacked any concept of the sound of a woodwind, what it would be like for one to have that sensation would then be correspondingly more generic. If one lacked even the concept of a sensation's being of a sound as against being of some other type of stimulus, having the sensation would for one be like merely having some indiscriminate sensory experience or other. This sequence makes it plausible that peeling away that weakest HOT would result, finally, in its no longer being like anything at all to have that sensation. Even though HOTs are just intentional states, and so have no qualitative properties, having HOTs does make the difference between whether there is or is not something it's like for one to have particular sensations.

Because HOTs seldom occur consciously, we cannot, from a first-person point of view, note the occurrence of HOTs when, and only when, we are in conscious sensory states. Still, the argument from wine tasting does draw on first-person considerations. We know in a first-person way that learning new concepts for sensory qualities is enough for us to come to be conscious of our sensory states as having those qualities. And on that basis, we can infer that nonconscious HOTs are responsible for there being something it's like for one to be conscious of our sensory states in that way. It's just that the direct correlation between nonconscious HOTs and conscious sensory states is unavailable from a first-person point of view.

Is it enough to have correlations inferred from first-person considerations? Or must we work completely within a first-person point of view if we are to show that HOTs are responsible for there being something it's like for one to be in conscious sensory states?

A theory of consciousness must explain the first-person aspects of our conscious states. But the explanation need not itself rely only on first-person aspects. Indeed, to demand otherwise is to make any such explanation viciously circular. So the factor responsible for there being something it's like to be in a sensory state need not itself be a first-person aspect of that state, nor even something available from a first-person point of view. The HOTs in virtue of which our mental states are conscious need not, themselves, be conscious thoughts.

Compare the causal relations conscious sensory states have to stimuli, behavior, and other mental states. These relations are typically unavailable from a first-person point of view; we must infer them from other considerations, both first- and third-person. Similarly, we may expect that whatever is responsible for there being something it's like for one to be in conscious sensory states is not directly accessible from a first-person point of view, but must instead be learned about by way of theoretical inference.

Some theorists have insisted that no correlations or theoretical developments could ever enable us to understand fully how physiological occurrences give rise to there being something it's like for one to be in conscious qualitative states. If so, perhaps we also cannot fully understand how HOTs could give rise to conscious qualities.

Joseph Levine calls this difficulty the 'explanatory gap' and argues that it results from our being able to conceive of physiological occurrences without conscious qualities. By contrast, he claims, it's inconceivable that water could boil at a different temperature, at least holding constant the rest of chemistry.[39] But our ability to understand things and the apparent limits on what we can conceive are always relative to prevailing theory, whether scientific or folk theory, as Levine's holding chemistry constant illustrates.

Since the appearance of an explanatory gap simply attests our current lack of a well-developed, suitable theory, theoretical advances pertaining to conscious qualitative states should substantially narrow whatever gap seems now to obtain. And, though we may never fully eliminate that gap, we seldom if ever have a complete understanding of how any commonsense, macroscopic phenomenon arises.[40]

The HOT model proceeds independently of physiology, but a similar explanatory gap

seems to arise, since we need to understand how nonconscious HOTs can result in conscious qualities. Causal connections are irrelevant here, since there need be no causal tie between a HOT and its target. Rather, HOTs result in conscious qualities because they make us *conscious of ourselves as being in certain qualitative states*, which results in the subjective impression of conscious mental qualities. And the considerations raised earlier in this section provide reason to hold that HOTs can actually do this.

V. Consciousness, Confabulation, and Function

In closing I turn briefly to two unexpected implications of the HOT hypothesis, indeed, of any theory on which a mental state's being conscious consists, as I've argued it must, in one's being transitively conscious of that state.

As we have seen, the HOTs in virtue of which mental states are conscious represent those states in more or less fine-grained ways. And the way our HOTs represent the states they are about influences what those states are like from a first-person point of view. What it's like for me to have a particular gustatory sensation of wine depends on how much detail and differentiation goes into the HOT in virtue of which that sensation is conscious. Given any particular sensory state, different HOTs would yield different ways it's like for one thing one to be in that state.

Since the HOT that accompanies any particular sensory state can be more or less fine-grained, it is not the sensory state alone that determines what HOT one will have. That will depend also on such additional factors as the size of one's repertoire of concepts, one's current interests, how attentive one is, and how experienced one is in making the relevant sensory discriminations.

This raises an interesting question. Since the sensation itself does not determine what HOT one has, why can't the HOT misrepresent the sensory state one is in? Why can't one be in a sensory state of one type, but have a HOT that represents one as being in a sensory state of some different sort? The HOT one has, moreover, determines what it's like for one to be in the relevant sensory state. So why wouldn't an erroneous HOT make it seem, from a first-person point of view, as though one were in a sensory state that one is not in fact in?

There is reason to believe that this actually happens. Dental patients sometimes seem, from a first-person point of view, to experience pain even when nerve damage or local anesthetic makes it indisputable that no such pain could be occurring. The usual hypothesis is that the patient experiences fear or anxiety along with vibration from the drill, and consciously reacts as though in pain. Explaining this to the patient typically results in a corresponding change in what it's like for the patient when drilling resumes, but the patient's sense of what the earlier experience was like generally remains unaltered. The prior, nonveridical appearance of pain is indistinguishable, subjectively, from the real thing.

Other striking examples occur in connection with our perceptual sensations. As Daniel Dennett notes in *Consciousness Explained*, parafoveal vision can produce only low-resolution sensations of most of the Marilyns in Warhol's famous painting,[41] but we are aware of them all as clear and focused. What it's like for us is a function not of the character of our sensations, but of how we're conscious of those sensations.

There is an also well-known tendency people have to confabulate being in various intentional states, often in ways that seem to make *ex post facto* sense of their behavior;[42] here it's plain that HOTs misrepresent the states that subjects are in. Similarly, it is very likely that repressed beliefs and desires are often actually conscious beliefs and desires whose content one radically misrepresents. Thus one might experience one's desire for some unacceptable thing as a desire for something else instead. In such a case, the desire is not literally unconscious; it is a conscious desire whose character is distorted by inaccurate HOTs. What it's like for one to have that desire fails accurately to reflect its actual content.[43]

The HOT hypothesis is not the only theory to make room for these things; any theory on which a mental state's being conscious consists in one's being transitively conscious of that state will do so. As long as a conscious state is distinct from one's transitive consciousness of it, the content of that transitive consciousness may misrepresent the state. Conscious states are states *we are conscious of ourselves as being in*, whether or not we are actually in them.

The idea that what it's like for one to be in a state is determined not by that state's intrinsic properties but by the way one's HOT represents it enables us to understand certain cases

that seem otherwise intractable to explanation. Suppose you're walking through the woods, stepping over branches as needed, but so deeply engrossed in conversation that you pay no conscious attention whatever to the branches. From a first-person point of view, you appear to have no thoughts about the branches; any thoughts about them you do have are not conscious thoughts.

To negotiate through the branches, however, you presumably need more than just thoughts about them; you must also have sensations of the branches. But from a first-person point of view, it may well also seem as though you have no such sensations. Unlike your thoughts, however, there is reason to doubt that your sensations of the branches literally fail to be conscious. It's not that there are no conscious sensations where one would expect sensations of branches to occur in one's visual field; the visual field does not seem to have gaps where the relevant sensations would be. Rather, the sensations that seem to you to be there are, roughly, just sensations of the undifferentiated rustic environment.

Why, then, are you unaware of your sensations of the branches? Plainly you have such sensations; that's how you manage to negotiate through the branches. And the sensations you have of the relevant part of the environment are all conscious; that's why your visual field doesn't seem to contain gaps. So it must be that the sensations are conscious not as sensations of branches, but only as sensations of the undiscriminated environment. We can explain this kind of occurrence only if the way one is transitively conscious of our sensations determines what it's like for one to have them. Compare Dennett's vivid example of looking straight at a thimble but failing to see it as a thimble. It's clear that one's sensation of the thimble is conscious, but one is conscious of it not as a sensation of a thimble but only, say, as a sensation of part of the clutter on a shelf.[44]

In the thimble and branches cases, what it's like for one to be in particular sensory states is informationally less rich than the states themselves. But the opposite also happens, as when we experience our low-resolution sensations of the parafoveal Marilyns as though they were clear and focused. The best explanation is that our HOTs about our blurry parafoveal sensations represent them as having high resolution; the way we are conscious of our sensations actually corrects them by, as it were, bringing them into focus and touching them up.[45] Indeed, this drives home the need to posit occurrent higher-order states, since the high-resolution information must be embodied in some occurrent state.

This disparity between the properties of our sensations and the way we're conscious of them has important implications. For an example, consider Wilfrid Sellars' well-known argument that the sensory qualities of sensations exhibit an 'ultimate homogeneity' that sets them apart from the particulate character of ordinary physical properties.[46] Sellars holds that this ultimate homogeneity derives from the way we conceive, in commonsense terms, of the perceptible properties of physical objects. Whatever the case about that, it is likely that those sensory qualities of sensations are themselves particulate. Being neurally based, the relevant sensory information will occur in the form of particular pixels that represent color, shape, motion, and the like. We experience such information, however, as ultimately homogeneous simply because that is how we are conscious of the relevant informational states. The way we are conscious of our sensations smooths them out, so to speak, and elides the details of their particulate, bit-map nature.

Dretske has noted that theories on which a state's being conscious consists in one's being transitively conscious of the state seem unable to explain how a mental state's being conscious could have any function.[47] Being transitively conscious of a state, on these theories, makes no difference to the state's nonrelational properties. So the state's being conscious will make no difference to its causal role nor, therefore, to its function.

It's easy to overestimate the degree to which a state's being conscious does actually play any role. It's inviting to think, for example, that a state's being conscious somehow enhances any planning or reasoning in which that state figures. But the role a state plays in planning and reasoning is due to the content the state has, and that content will be invariant whether or not the state is conscious. So whether or not a state is conscious will not affect the state's role in planning and reasoning. We find it tempting to insist that a state's being conscious affects planning and reasoning when we consider actual cases in which the planning and reasoning are conscious. But those cases tell us nothing unless we compare them to nonconscious cases, to which we have no first-person access. Intuitions cannot help here.

In any event, Dretske has misdescribed the situation. On the HOT hypothesis, a conscious

state is a compound state, consisting of the state one is conscious of together with a HOT. So the causal role a conscious state plays is actually the interaction of two causal roles: that played by the state itself and that played by the HOT.[48] This explains how a state's being conscious may to some extent matter to its causal role. Moreover, the way one is conscious of a conscious state may not fully match the target state one is actually in. In those cases, the causal role played by the HOT will matter even more. State consciousness does, after all, make some small difference to the function mental states have.[49]

But what, then, of the compelling intuition that a mental state's being conscious does make a large and significant difference to its mental functioning? That intuition is very likely due to the sense we have that our conscious thoughts, desires, and intentions occur freely and that this apparent freedom enhances our ability to reason and make rational choices. But our sense that these states occur freely itself arguably results from the way we are conscious of those states. Because we are seldom if ever conscious of anything as causing our conscious thoughts and desires, we have the subjective impression that they are uncaused, and hence free. So it seems that just being conscious of these states makes a significant difference to the role they can play in our lives. It is because the way we are conscious of our intentional states presents them as free and uncaused that their being conscious seems to matter to our ability to reason and make rational choices.

I have argued that the HOT hypothesis explains how conscious states differ from nonconscious mental states, and why, to the extent that it does, state consciousness has a function. Moreover, the hypothesis squares well with there being something it's like to be in conscious sensory states. We can provisionally conclude that the hypothesis deals satisfactorily with the phenomenon of state consciousness, even for the special case of sensory states.

NOTES

1. *Pace* John R. Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992); see note 8, below.

   I use 'intentional state' here to refer to states, like beliefs and desires, that exhibit propositional content along with some mental attitude.
2. For some related observations about different uses of 'consciousness.' see Edmund Husserl, *Logical Investigations* (London: Routledge & Kegan Paul, 1970), II, pp. 535–36.
3. 'Ordinary' is to exclude so-called hypnogogic dreams, which occur in a semi-waking state.

   Intuitions here are in any case hardly decisive. Are very vivid dream states conscious states? Must we be conscious when we're in them? Since it's far from clear what to say about these matters, it may well be that conscious states can occur without the creature itself being conscious.
4. There is, of course, nothing it's *like* for such a creature to be conscious—nothing it's like *for* the creature. But that doesn't mean there's nothing it *is* to be conscious.
5. Even if all sensations were conscious, what it is for a sensation to be *of* something would be a function not of its being conscious, but rather of the ways it qualitatively resembles and differs from other comparable sensations.

   Strictly speaking, mental states aren't conscious of things; rather, it's creatures that are conscious of things in virtue of their being in mental states.
6. The best-known version of the example is due to D. M. Armstrong, "What Is Consciousness?" in his *The Nature of Mind* (St. Lucia, AUS: University of Queensland Press, 1980), 55–67, p. 59. See Dretske's *Naturalizing the Mind* (Cambridge, MA: MIT Press/Bradford Books, 1995), pp. 104–5.
7. "Conscious Experience," *Mind* 102, no. 406 (April 1993): pp. 263–83; reprinted in Dretske, *Perception, Knowledge, and Belief: Selected Essays* (Cambridge, UK: Cambridge University Press, 2000), 113–37, p. 123; *Naturalizing the Mind*, chapter 4.
8. "Conscious Experience," pp. 125–28; cf. *Naturalizing the Mind*, pp. 112–3.
9. One might object that we are, in any case, conscious of our conscious states when we are introspectively aware of them. To forestall this objection, Dretske has recently argued that introspection resembles what he calls displaced perception. Just as we come to know how full the gas tank is by looking at the gauge, so we come to know what mental state we're in by noticing what we're seeing. We thereby come to be conscious *that* we're in some particular mental state, but not conscious *of* that state. (Dretske, "Introspection," *Proceedings of the Aristotelian Society*, CXV [1994/95]: pp. 263–78, and *Naturalizing the Mind*, chapter 2.)

   On this ingenious proposal, introspection is a matter of coming to know how one represents things (274–75). But introspection is better construed as knowing what mental state one is in, independently of how that state represents nonmental reality. But even if Dretske's right about what introspection is, just seeing that I represent things as being a certain way won't yield introspection unless I see this consciously. So either the argument rests on Dretske's assumption that all mental states are conscious, or he must give a different account of what it is for states to be conscious.
10. "Conscious Experience," p. 128; cf. pp. 117–18.
11. In his useful "Dretske on HOT Theories of Consciousness," William Seager independently

gives a similar account of how Dretske's argument fails to undermine the HOT hypothesis (*Analysis* 54, no. 1 [January 1994]: pp. 270–76, especially pp. 275–76).

12. 'But readers who were only thing-aware of the difference between Alpha and Beta [the two arrays in Dretske's example] were not fact-conscious of the difference between Alpha and Beta.' ("Conscious Experience," p. 128.)
13. John R. Searle also denies that we are conscious of our conscious mental states, though for reasons different from Dretske's. '[W]here conscious subjectivity is concerned, there is no distinction between the observation and the thing observed' (*The Rediscovery of the Mind*, p. 97). The context makes clear that Searle is denying not just that we can observe our conscious states, but that we are conscious of them at all, in the way we're conscious of other things: 'We cannot get at the reality of consciousness in the way that, using consciousness, we can get at the reality of other phenomena' (pp. 96–97). This is because 'where conscious subjectivity is concerned, there is no distinction between the observation and the thing observed' (p. 97).

    Searle argues for this by appeal to the idea that we can describe consciousness only in terms of what it's consciousness *of* (p. 96). But even if that's so, there will be states in virtue of which we are conscious of things. So it doesn't follow that there aren't states in virtue of which we are conscious of our conscious states.
14. '[T]he word 'thought' applies to all that exists in us in such a way that we are immediately conscious of it' (Geometrical Exposition of the *Second Replies, Oeuvres de Descartes*, Charles Adam and Paul Tannery, eds., [Paris: J. Vrin, 1964–75], VII, p. 160).
15. Nonconscious mediation, moreover, might well occur; factors of which we're in no way conscious often causally mediate among distinct mental states, even when we're aware of them from a first-person point of view.

    Our intuitive sense that we're not conscious of our conscious states in any way that's mediated may be what leads Searle to claim that there's no way in which we're conscious *of* our conscious states (see note 13). It also distinguishes this case from the way we're perceptually conscious of things, in which we are sometimes conscious of the intervening medium.
16. A slight adjustment to this is needed. One might hold a theory on which an inference mediates between our being conscious of our conscious states and the states themselves, though we're conscious of that inference only by another inference based on the theory. (I thank Eric Lormand for raising this possibility.) We would still count as conscious the same states, even though the theory makes us conscious of the inferential mediations. We can provide for this by stipulating that if a state is conscious, we're conscious of it in a way that does not require that we be conscious of any inference that may occur. Our being conscious of the state may rely on some inference, but not on our being conscious of it.

    This handles a related possibility as well. Suppose that inferences of which we're not conscious normally mediate between our being conscious of our conscious states and those states. Even if we somehow became conscious of some of those inferences without benefit of theory, we'd count the same states as conscious. The adjusted stipulation provides for this. Since nothing in what follows hinges on this sort of thing, I'll omit this qualification.

    If the way we're conscious of our conscious states were sometimes based on conscious inference, we'd then know how we come to be conscious of those states. Though we're conscious of our conscious states, we generally don't, from a first-person point of view, have any idea how we come to be conscious of them. That ignorance helps explain the air of mystery that surrounds state consciousness.
17. The inference that consciously mediates between mental states and one's being conscious of them need not begin with the mental state to one's being conscious of it; typically, the conscious inference would start, instead, from noticing one's behavior or from the remarks of others. And because those things are causally due to one's mental state, such an inference counts as mediating between a mental state and one's being conscious of it.

    We need not independently preclude reliance on observation. Intuitively, one's being conscious of a mental state can be immediate even if it relies on observation, so long as one is not aware of its doing so. And that will be so if there's no reliance on any conscious inference. E.g., if one observes one's happy gait and so, without any inference of which one is aware, takes oneself to be happy, the way one is conscious of being happy is intuitively immediate.
18. Although we recognize on reflection that mediation does in fact occur, no conscious inference normally mediates, and as we've seen, that's what matters for the intuition of immediacy.
19. On this, see David M. Rosenthal, "The Colors and Shapes of Visual Experiences," in *Consciousness and Intentionality: Models and Modalities of Attribution*, Denis Fisette, ed., (Dordrecht: Kluwer Academic Publishers, 1998), pp. 137–69; and "The Independence of Consciousness and Sensory Quality," in *Consciousness: Philosophical Issues, 1, 1991*, Enrique Villanueva, ed., (Atascadero, CA: Ridgeview Publishing Co., 1991), pp. 15–36, reprinted in *Consciousness and Mind* (Oxford, UK: Oxford University Press, 2005).
20. These considerations are reminiscent of an argument of Aristotle's at *de Anima* III, 2, 425b12–14, though Aristotle also held that the redness of our perceptions is the very same quality as the redness of physical objects (e.g., *de Anima* II, 5, 418a4; II, 11, 423b31; II, 12, 424a18; III, 2, 425b23).

    Perhaps the qualities of the higher-order states are those our sensory states seem to have, and the lower-order qualities do not figure in what it's like for us to be in sensory states. But locating the qualities that figure in what it's like to be in sensory states at the higher level doesn't help explain the qualitative dimension of those states.
21. The concern that nonlinguistic creatures can't be in intentional states with such sophisticated content may also motivate preference for the perceptual model, since perceiving is a less sophisticated mental phenomenon. But little conceptual richness is needed to be in such intentional states. The concept of self, e.g., need involve no more than the distinction between oneself and everything else. And the state itself can be conceptualized in a relatively minimal way, say, just as some way the creature is.

22. So HOTs are not simply about intentional contents, but about full-fledged intentional states: contents plus mental attitudes.

    If I doubt or wonder whether a particular physical object is red, I'm conscious of that object; similarly if I expect, hope, or desire that it is. But it's not the doubt, wonder, hope, or desire that makes me conscious of the object. Rather, if I doubt whether the object is red or desire or suspect that it is, I must also think assertorically that the object is there, or exists, and I'm conscious of the object in virtue of my having that assertoric thought. This is evident because, in such a case, I wouldn't be conscious of the object as red, but just as something that exists. The content of my consciousness is determined not by the content of my nonassertoric intentional state, but by the assertoric state. Similarly with intentional states about our own mental states; being in nonassertoric intentional states about one's mental states make one conscious of being in those states only if they require one also to have the assertoric thought that one is in that state.

    It's worth noting an argument of Robert M. Gordon that many emotions must be accompanied by corresponding beliefs; being angry that *p*, e.g., requires believing that *p*. (*The Structure of Emotions: Investigations in Cognitive Philosophy* [Cambridge, UK: Cambridge University Press, 1987], pp. 47ff). If so, the required belief would explain why, when one's angry that *p*, one is conscious of whatever ⁅*p*⁆ is about. In any case, this result depends on describing the emotion in terms of its intentional content. Thus, if one describes a person not as being angry that *p*, e.g., but as being angry because *p*, no corresponding belief is implied.

23. This helps deal with an interesting objection. Freudian theory may seem to posit states that are nonconscious despite their being accompanied by suitable HOTs. (This idea has been pressed by Georges Rey and Stephen Schiffer.) But it's not easy to come up with convincing examples. Pleasure or guilt about repressed states won't do because pleasure and guilt aren't assertoric; so we often aren't conscious of the objects of our pleasure or guilt—even when those states are conscious.

    Even if we could come up with plausible examples, moreover, it is far from obvious that Freudian theory requires that we describe the situation as involving nonconscious states accompanied by HOTs, since there typically are several equally good explanations for any such phenomenon. It's also important to note that repressed states are seldom nonconscious states. Rather, they're typically states we disguise by radically misrepresenting their content, or distract ourselves from by creating elaborate mental noise. See p. 29, below.

24. According to Searle, the intentional content of perceptual states always refers to those very states; if I see a yellow station wagon, the content of my visual perception is 'that there is a yellow station wagon there and that there is a yellow station wagon there is causing this visual experience' (*Intentionality: An Essay in the Philosophy of Mind* [Cambridge, UK: Cambridge University Press, 1983], p. 48). If the content of every perceptual state were partly that one is in that state, then on the HOT hypothesis, just being in the state would make one conscious of it, and nonconscious perceptions would be impossible. (I am grateful to Gilbert Harman for raising this concern.) Moreover, perceiving something does presumably make one conscious of that thing, arguably because the mental attitude of perceiving is assertoric.

    Searle's argument for this claim appeals to the truth conditions of perceptions; a state's intentional content 'determines under what conditions it is satisfied' (p. 48), and one perceives a thing only if it causes one's perception. But the conditions under which the perception is satisfied are simply that there's a yellow station wagon there, not also that the perception is caused by there being a yellow station wagon there. The causal condition is relevant not to the truth of what I perceive, but of whether I perceive it.

    These considerations do, however, point toward an explanation of how many perceptual states do come to be conscious. We assume as a general belief about the world that the states of affairs we perceive normally cause the relevant perceptual states. When one has the (typically nonconscious) thought that a perceived state of affairs has caused the perceptual state, that thought results in a HOT that one is in the perceptual state, and thus results in that state's being conscious.

25. See, e.g., Dretske, "Conscious Experience," especially Section 4; also Ned Block, "Review of Daniel C. Dennett, *Consciousness Explained*," *The Journal of Philosophy* XC, no. 4 (April 1993): pp. 181–93, who alludes on p. 182 to the HOT hypothesis.

26. This argument is developed in detail in my "Thinking That One Thinks," in *Consciousness: Psychological and Philosophical Essays*, Martin Davies and Glyn W. Humphreys, eds., (Oxford: Basil Blackwell, 1993), pp. 197–223. On the connection between thought and genuine speech, see my "Intentionality," *Midwest Studies in Philosophy*, X (1986): pp. 151–84. Both will be reprinted in *Consciousness and Mind* (Oxford, UK: Oxford University Press, 2005).

    The argument relies on creatures that can describe their mental states. But noninferential reportability simply helps fix the extension of 'conscious state'; many nonlinguistic creatures are also in conscious states.

    Special issues arise about qualitative states, since there is no such thing as verbally expressing a perceptual sensation. We can express perceptions, but only because perceptions, unlike sensations, have an intentional aspect and it's that intentional component that we can verbally express. The same may also hold for bodily sensations; though we can express a pain by uttering 'ouch,' it's unclear that 'ouch' counts as a verbal, as opposed to nonverbal, form of expressing. And, though saying 'It hurts' is linguistic, that reports the pain, rather than expressing it. Still, creatures with suitable linguistic ability can noninferentially report their conscious states, whether the states are intentional or sensory.

    These considerations have a bearing on the perceptual model. When a state is conscious, creatures with the relevant linguistic ability can express their transitive consciousness of the state. If there were a higher-order perception of the state, one's report would verbally express only the intentional

component of that higher-order perception. But that's in effect just to express a HOT. So the argument from reporting and expressing shows that if the transitive consciousness of a conscious state did have a sensory aspect, that sensory aspect would be irrelevant to the state's being intransitively conscious.

27. Dennett and Harman have independently pressed this reply in conversation, and it receives tacit expression in Dennett's view that '[c]onsciousness is cerebral celebrity' ("The Message Is: There is no *Me*dium," *Philosophy and Phenomenological Research* LIII, no. 4 [December 1993]: 919–31, p. 929). See also Dennett, *Consciousness Explained*, chapter 10 and especially p. 315.
28. Peter Carruthers, *Language, Thought, and Consciousness: An Essay in Philosophical Psychology* (Cambridge, UK: Cambridge University Press, 1996); and *Phenomenal Consciousness: A Naturalistic Theory* (Cambridge: Cambridge University Press, 2000).
29. See Lawrence Weiskrantz, *Blindsight* (Oxford: Oxford University Press, 1986); and *Consciousness Lost and Found: A Neuropsychological Exploration* (Oxford: Oxford University Press, 1997).

    There is reason to think that discrimination of stimuli with different form may be due to discrimination of orientation, rather than of form itself (*Blindsight*, p. 84). Van Gulick has argued that this shows that blindsight does not involve states with phenomenal properties like those of conscious visual sensations. ("Deficit Studies and the Function of Phenomenal Consciousness," in *Philosophical Psychopathology*, George Graham and G. Lynn Stephens, eds., [Cambridge, MA: MIT Press, 1994].) But that conclusion follows only if one assumes that sensory qualities must be integrated in just the way they are in normal conscious cases.
30. A classical example is the so-called cocktail-party effect. We typically screen out the sounds of conversations other than our own, though mention of one's name in a screened-out conversation normally causes one's attention suddenly to shift to that conversation.
31. Compare parallel arguments that certain nonconscious states have mental properties because of the roles they play in mental processes; e.g., J. A. Fodor, "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology," *The Behavioral and Brain Sciences* III, no. 1 (March 1980): pp. 63–73.
32. See Thomas Nagel's "What is it like to be a bat?" *The Philosophical Review* LXXXIII, no. 4 (October 1974): pp. 435–50; "Panpsychism," in *Mortal Questions* (Cambridge, UK: Cambridge University Press, 1979), pp. 181–95; and *The View from Nowhere* (New York: Oxford University Press, 1986), chapters 1–4.
33. "On a Confusion about a Function of Consciousness," *The Behavioral and Brain Sciences*, 18, no. 2 (June 1995): 227–47, p. 231; emphasis Block's. See also Block, "Review of Dennett's *Consciousness Explained,*" p. 184; "Begging the Question against Phenomenal Consciousness," *The Behavioral and Brain Sciences* 15, no. 2 (June 1992): pp. 205–6; "Consciousness and Accessibility," *The Behavioral and Brain Sciences* XIII, no. 4 (December 1990): pp. 596–98.
34. It's not that the states we report are nonconscious, but nonconscious states influence what we report and how we do it.
35. Block's definition of access consciousness in terms of a state's being 'poised' for certain things gives a dispositional mark of such consciousness. (In the review of Dennett's book, he uses the phrase 'freely available' [p. 182].) That is compatible with access consciousness's consisting in a subject's being transitively conscious of a mental state, rather than simply being disposed to be conscious of it. States we are transitively conscious of have many dispositional properties, among them being reportable and introspectible.
36. Block distinguishes a third concept of consciousness, which he calls reflective consciousness (Review of Dennett's *Consciousness Explained*," p. 182) or monitoring consciousness ("On a Confusion about a Function of Consciousness ," p. 235). According to Block, a state is conscious in this way if one has a HOT about it. But the states he counts as reflectively or monitoring conscious are states that we're introspectively conscious of: states that we're conscious of being conscious of. This is a distinct notion of consciousness, but Block is mistaken to define it in terms of having HOTs. Rather a state has monitoring consciousness, in his terms, only if one has a *conscious* HOT about it. See note 25.

    For more on Block, see Rosenthal, "Phenomenal Consciousness and What It's Like," *The Behavioral and Brain Sciences*, 20, no. 1 (March 1997) pp. 64–65; "The Kinds of Consciousness," MS, and "How Many Kinds of Consciousness," MS.
37. Of course, the relevant sensory states will often have been conscious before one acquired the more fine-grained concepts, but conscious only with respect to qualities individuated in a more course-grained way. E.g., one might initially be conscious of a particular type of olfactory sensation solely as being winelike, and subsequently become conscious of it in terms of more fine-grained sensory qualities.
38. One might argue that the new concepts pertain not to the distinguishing properties of our conscious sensory experiences, but rather to the perceptible properties of the perceived physical objects and processes, e.g., the wine or the musical performance. (See Harman, "The Intrinsic Quality of Experience," *Philosophical Perspectives, 4: Action Theory and Philosophy of Mind* [1990]: pp. 31–52.) But it's clear that in the cases just imagined we also focus introspectively on the distinguishing properties of our conscious sensory states. So those cases involve new concepts of the distinguishing properties of sensory states.
39. "On Leaving Out What It's Like," in *Consciousness: Psychological and Philosophical Essays*, Martin Davies and Glyn W. Humphreys, eds., (Oxford: Basil Blackwell, 1993), 121–36, p. 134; and *Purple Haze: The Puzzle of Consciousness* (New York: Oxford University Press, 2001), p. 79. See also "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* LXIV, no. 4 (October 1983): pp. 354–61. For related arguments, see David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996).

Similarly, Nagel claims we have a purely rational understanding of why 'heat caus[es] water to boil, rocks caus[e] glass to break, magnets induc[e] electric current, [and] the wind mak[es] waves' ("Panpsychism," p. 186), but currently lack any understanding of how physical heat, e.g., or a brain process, could causally necessitate a pain or other sensation ("Panpsychism," p. 187).

40. See my "Reductionism and Knowledge" in *How Many Questions?* Leigh S. Cauman, Isaac Levi, Charles Parsons, and Robert Schwartz, eds., (Indianapolis: Hackett Publishing Co., 1983), pp. 276–300.
41. Daniel C. Dennett, *Consciousness Explained* (Boston: Little, Brown and Co., 1991), p. 354. See pp. 53–54 for Dennett's striking illustration of these limits in attempting to discern the color of playing cards seen parafoveally at arm's length.
42. For a classic study, see Richard E. Nisbett and Timothy DeCamp Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* LXXXIV, no. 3 (May 1977): pp. 231–59. Nisbett and Wilson's influential study focused not only on cases in which subjects confabulate stories about the causes of their being in particular cognitive states, but also on cases in which they confabulate accounts about what states they're actually in.
43. Perhaps such erroneous HOTs might figure also in apparent self-deception.

    It may sometimes be difficult to tell whether a HOT misrepresents an actual target or the HOT has only a notional target, and there is an actual state that simply isn't conscious. Indeed, it may well be arbitrary within a certain range of cases which way we describe a case.
44. *Consciousness Explained*, p. 336.

    Similarly, in the cocktail-party effect, one's attention shifts to a previously unattended conversation in which one's name was mentioned. So one must have been hearing the articulated words in that conversation, though to consciousness it seemed just to be background din.

    Robust experimental findings, e.g., those involving masked priming, also provide compelling evidence that what it's like to have a sensation sometimes diverges from the properties of the sensation itself. In masked priming, subjects report being unaware of qualitative input whose presence is evident from its effect on subsequent cognitive behavior. For a classic study, see Anthony J. Marcel, "Conscious and Unconscious Perception: Experiments on Visual Masking and Word Recognition," *Cognitive Psychology* 15 (1983): pp. 197–237.

    Experimental work on change blindness also provides vivid evidence for divergence of how we're conscious of our sensations from their actual properties. Subjects here fail consciously to register visible changes so salient that it's overwhelmingly likely that corresponding changes do occur in their visual sensations. So subjects' sensations diverge from how they're aware of them. Moreover, the compelling impression we all have of being continuously conscious of salient qualitative detail is evidently erroneous. See John Grimes, "On the Failure to Detect Changes in Scenes across Saccades," in *Perception*, Kathleen Akins, ed., (New York: Oxford University Press, 1996), pp. 89–110; Daniel J. Simons, "Current Approaches to Change Blindness," *Visual Cognition* 7 (2000): pp. 1–16; and Ronald A. Rensink, "The Dynamic Representation of Scenes," *Visual Cognition*, 7, no. 1/2/3 (January 2000): pp. 17–42; and "Seeing, Sensing, and Scrutinizing," *Vision Research*, 40, no. 10–12 (2000): pp. 1469–87.
45. In aesthetic experience, also, how we are conscious of a sensation presumably outstrips that sensation's qualitative character.

    For more on sensations' diverging from the way we are conscious of them and the way HOTs function in that connection, see "Sensory Qualities, Consciousness, and Perception," in *Consciousness and Mind*, and "Consciousness and Metacognition" in *Metarepresentation: Proceedings of the Tenth Vancouver Cognitive Science Conference*, Daniel Sperber, ed., (New York: Oxford University Press, 2000), pp. 265–95.
46. Often referred to as Sellars' 'grain argument.' Wilfrid Sellars, "Philosophy and the Scientific Image of Man," in *Frontiers of Science and Philosophy*, Robert G. Colodny, ed., (Pittsburgh: University of Pittsburgh Press, 1962), pp. 35–78; reprinted in *Science, Perception and Reality*, 1–40, p. 36; also p. 35, and "Phenomenalism," also in *Science, Perception and Reality*, 60–105, pp. 103–5.

    Cf. Peter Carruthers' claim that 'perceptual information is *analogue* (that is, 'filled in' and continuous),' and 'the subjective aspect of an experience just *is* analogue information about [physical] red, presented to a cognitive apparatus having the power classify states as information carriers, as well as to classify the information carried' (Peter Carruthers, *Language, Thought, and Consciousness: An Essay in Philosophical Psychology* [Cambridge, UK: Cambridge University Press, 1996], pp. 167, 214).
47. *Naturalizing the Mind*, p. 117.
48. The interaction of the two roles may not be additive; the causal properties of the HOT may interact with those of the state in such a way that the original causal properties of the state are modified, or even blocked altogether.
49. So conscious inessentialism, on which every intelligent activity we perform consciously could be performed without its being conscious, is mistaken. The label is due to Owen Flanagan, who rejects the thesis (*Consciousness Reconsidered* [Cambridge, MA: Bradford Books/MIT Press, 1992], pp. 5, 129ff.).

Visual Qualia and Visual Content Revisited

Michael Tye

Experiences vary widely. For example, I run my fingers over sandpaper, smell a skunk, feel a sharp pain in my finger, seem to see bright purple, become extremely angry. In each of these cases, I am the subject of a mental state with a very distinctive subjective character. There is something it is *like* for me to undergo each state, some phenomenology that it has. Philosophers often use the term 'qualia' to refer to the introspectively accessible properties of experiences that characterize what it is like to have them. In this standard, broad sense of the term, it is very difficult to deny that there are qualia. There is another, more restricted use of the term 'qualia,' under which qualia are intrinsic, introspectively accessible, nonrepresentational qualities of experiences. In my view, there are no qualia, conceived of in this way. They are a philosophical myth.

Elsewhere (Tye 1995, 2000) I have argued that all experiences have representational content and that what it is like to undergo an experience is a matter of a certain sort of representational content the experience has. This view has come to be known as representationalism. In this paper, I shall not try to defend representationalism generally. My concern here is exclusively with the case of visual experience and visual qualia. I shall try to show not only that no good reasons have been adduced for believing in visual qualia in the second sense (hereafter Qualia) but also that, upon proper reflection, the most natural view, is that there are none.

The paper is divided into three sections. In Section I, I say some more about what sorts of properties of visual experience visual Qualia are supposed to be and what it is that I am committed to in denying their existence. In Section II, I discuss a variety of arguments and examples that purport to show that there are visual Qualia. Finally, in Section III, I make some brief comments about representationalism with respect to visual experience and also about the overall significance of my attempt to account for visual experience without Qualia.

I

Consider a painting of a tiger. Viewers of the painting can apprehend not only its content (i.e., its representing a tiger) but also the colors, shapes, and spatial relations obtaining among blobs of paint partly by virtue of which it has that content. It has sometimes been supposed that being aware or conscious of a visual experience is like viewing an inner picture. So, for example, on this conception of vision, if I train my eyes on a tiger in good light, I am subject to a mental picture-like representation of a tiger, introspection of which reveals to me both its content and its intrinsic, nonintentional features partly by virtue of which it has that content. These intrinsic, nonintentional features are not literally colors and shapes of parts of my mental quasi-picture, as in the case of a real picture. After all, it would obviously be absurd to suppose that parts of my brain are orange and black striped when I see a tiger, and it is surely no less absurd to suppose that parts of my soul are. So, whether visual experiences are physical or not, even on a pictorial conception, their introspectible, intrinsic properties are not colors and shapes.

Anyone who believes that there are visual Qualia must at least believe that visual experiences are like pictures to the extent that they (or a mental object they involve) have intrinsic, nonintentional or nonrepresentational features which are accessible to introspection and partly by virtue of which the experiences represent what they do.[1] It is Qualia, so conceived, that are my target. In opposing Qualia, I am not denying, of course, that the contents of visual experiences are introspectible. Nor am I denying that visual experiences have intrinsic, nonintentional or nonrepresentational features. If, as is widely believed, visual experiences are neural items, they will certainly have intrinsic physico-chemical properties.

To emphasize: the rejection of visual Qualia is *not* tantamount to a rejection of the view that there is something it is like for the subjects of

Revised from "Visual qualia and visual content," in T. Crane, ed., *The Contents of Experience* (Cambridge University Press, 1993), with the permission of Cambridge University Press.

visual experiences. On the contrary, as I noted in the introduction, the view I accept is that what it is like to have a visual experience (what is sometimes called 'the phenomenal character of the experience') is a matter of a certain sort of representational content that the experience has.[2] A consequence of this position is that necessarily any two visual experiences that are exactly alike in their representational contents are exactly alike in their phenomenal character. To refute my position, it suffices to specify a clear counter-example to this generalization. I know of no such counter-example. In the next section, I shall consider a variety of putative counter-examples together with a number of other objections.

II

The Argument from Introspection

Standing on the beach in Santa Barbara a number of summers ago on a bright sunny day, I found myself transfixed by the intense blue of the Pacific Ocean. Was I not here delighting in the phenomenal aspects of my visual experience? And if I was, doesn't this show that there are visual Qualia?

I am not convinced. It seems to me that what I found so pleasing in the above instance, what I was focusing on, as it were, were a certain shade and intensity of the colour blue. I experienced blue as a property of the ocean not as a property of my experience. My experience itself certainly wasn't blue. Rather it was an experience that represented the ocean as blue. What I was really delighting in, then, was a quality *represented* by the experience, not a quality *of* the experience. It was the color, blue, not anything else that was immediately accessible to my consciousness and that I found so pleasing. This point, I might note, seems to be the sort of thing G. E. Moore had in mind when he remarked that the sensation of blue is diaphanous (see Moore 1922, p. 22). When one tries to focus on it in introspection one cannot help but see right through it so that what one actually ends up attending to is the real colour blue.

There is another rather different way in which a straightforward appeal to introspection might be made on behalf of Qualia. The visual experience I had that day in Santa Barbara, as I stood entranced by the colour of the sea, was, to my consciousness, very similar to a colour photograph I might have taken of the same scene. My experience, then, was a picture-like representation of the sea, and my awareness of it was something like my viewing a picture. Since, as I noted in the last section, pictures evidently have accessible intrinsic qualities partly by virtue of which they represent the world, so too, by analogy, do visual experiences.

The most obvious problem with this appeal is that it is not at all clear that my visual experience, while viewing the ocean, was *really* similar to a colour photograph of the ocean. The only undeniable similarity here is between my experience and the experience I would have undergone had I viewed an appropriate photograph. The fact that these experiences are similar shows nothing about the way in which their contents are encoded. What I deny, then, is that the *format* of visual representations—the way in which they encode their contents—is given in introspection. What introspection reveals are simply aspects of the contents themselves.

The second objection I have is simply that even if visual experiences are, in an important sense, picture-like, it evidently does not follow that they have Qualia. One could hold, e.g., that visual experiences have intrinsic qualities partly by virtue of which they represent while denying that these qualities are introspectively accessible (see Harman 1990). Such a position still permits the possibility that visual experiences are picture-like, e.g., with respect to the representation of spatial relations.[3] But it leaves no room for Qualia.

The Argument from Hallucination

Suppose that Paul hallucinates a pink square object. Then there is something that Paul hallucinates. But what Paul hallucinates is not a real pink, square, physical object—Paul, after all, is hallucinating not seeing. So what Paul hallucinates must be a mental object, an idea or an appearance. Now mental objects are not literally coloured nor do they literally have shape. So the terms 'pink' and 'square' in application to what Paul hallucinates must pick out special properties of which Paul is directly aware. These properties are Qualia. Since seeing can be indistinguishable from hallucinating, such properties are present in cases of veridical perception too.

I lack the space to comment on all that is wrong with this argument. When Paul hallucinates in the above case he has an experience *of* a pink square object. *There is*, then, a definite content to Paul's hallucinatory experience. But there is no object, mental or otherwise, that Paul

hallucinates. Furthermore the fact that Paul's experience has a certain content no more requires that there really be a pink square object than a picture's representing a three-headed monster, say, requires that there really be any monsters.

Consider the following parallel. Paul wants a blue emerald to give to his wife. There are no blue emeralds. It does not follow that Paul wants the idea of a blue emerald to give to his wife. That he already has. What he wants is that his wife be given a blue emerald (by him). His desire, then, is the desire it is in virtue of its having specific content. When Paul reflects upon or introspects his desire what he is aware of is this content rather than any peculiar qualities of a special mental particular upon which his desire is directed. Likewise when Paul hallucinates a pink square what he introspects, I maintain, is the content of his hallucinatory experience. The qualities of which he is introspectively aware enter into this content (color and shape qualities, among others), and, given that Paul is hallucinating, they belong to nothing before Paul at all. This, it seems to me, is the common-sense view. The idea that the terms 'pink' and 'square' in the context 'Paul hallucinates a pink square' stand for special, phenomenal qualities of which Paul is aware and hence have entirely different meanings from those they have in, say, 'The piece of glass is pink and square' is, on the face of it, very strange indeed. The argument from hallucination does nothing to make this idea palatable.

Visual Qualia without Visual Content

Here is a related argument. Suppose you look at a bright light and turn away. You have an after-image that is red and round, say. In this case you are subject to a visual experience but your experience has no representational content. What it is like for you, then, cannot be a matter of the content of your experience. Rather it must be due to visual Qualia[4] (see Jackson 1977).

It seems to me no more plausible to take the terms 'red' and 'round,' as they apply to an after image, as denoting intrinsic qualities of the image than it is to take the terms 'loud' and 'high-pitched,' as they are employed in connection with the graphical representations of sounds, as denoting intrinsic qualities of oscilloscope readings. People who work with such readings frequently use terms like 'loud' and 'high-pitched' in application to the readings themselves. (This example is due to Ned Block 1983, pp. 516–17.) It is obvious that in this usage what the terms really pick out are features of sounds *represented* by the readings (loudness and high pitch respectively). Analogously it seems to me that what the terms 'red' and 'round' signify, in application to an after-image, are properties represented by the after-image experience. In my view, there is no after-image that is the mental object of the experience. One who has a red, round after-image is subject to a visual experience, produced by looking at a bright light, the representational content of which is (very crudely) that something red, round and filmy is hovering in space. Since there is no such thing hovering in space, the experience is illusory. There is, then, I claim, a definite content to the visual experience after all.

The fuzziness of most after-images, I might add, is most easily accounted for by supposing that it is a straightforward reflection of the representational impoverishment of the relevant visual experiences. If I have a red, round, fuzzy after-image, my experience does not 'say' exactly where the boundaries of the nonexistent red, round thing lie.

Experienceless 'Sight'

Albert is a very remarkable man. He is blind and he has been so since birth. Nevertheless when he faces objects and concentrates fiercely, thoughts pop into his head—he knows not where they come from—about the visual properties and relations of the objects. These thoughts are so detailed that *content-wise* they are just as rich as the visual experiences sighted people have in the same circumstances. Indeed were one to pay attention merely to the contents of Albert's thoughts, as expressed in his verbal descriptions of what is before him, one would be convinced that he is seeing. But Albert has no visual experiences. For Albert there is experientially no difference between his thoughts on such occasions and his thoughts when he ruminates on mathematics or art or life in general. In each case thoughts just occur and he is introspectively aware of no more than the contents of his thoughts. There is, then, an enormous felt difference between Albert and his sighted fellows at the times at which Albert seems to be seeing. This difference is one that Albert himself would come to appreciate in detail were he to gain sight. It is a difference that can only be explained on the assumption that Albert's inner states lack visual Qualia.[5]

Not so. There is another explanation. Intuitively, the content of visual experience goes far beyond any concepts the subject of the experience may have. Consider, for example, my visual experience of a determinate color hue—red_{19}, say. This does not demand possession of the concept, red_{19}. For I certainly cannot recognize that hue as such when it comes again. I cannot later reliably pick it out from other closely related hues. My ordinary color judgements, of necessity, abstract away from the myriad of details in my experiences of color. The reason presumably is that without some constraints on what can be cognitively extracted, there would be information overload.

Likewise, the representation of viewpoint-relative shape properties is naturally taken to be nonconceptual in some cases. Presented with an unusual shape, I will have an experience of that shape, as seen from my viewpoint. But I need have no concept for the presented shape. I need have no ability to recognize that particular viewer-relative shape when I experience it again. Arguably, even the representation of viewpoint-independent shapes is sometimes nonconceptual.[6] But clearly some representation in visual experience is a conceptual matter (e.g., the representation of object types such as car, ball and telescope).

Some seek to explain the richness of visual experience conceptually by noting that even though the subject often has no appropriate non-indexical concept, he or she is at least aware of the pertinent feature, e.g., red_{19}, as *that* color or *that* shade or *that* shade of red (McDowell 1994). This seems to me unsatisfactory. Intuitively, one can have a visual experience without having such general concepts as *color, shade*, or *shade of red.* Indeed, one can have a visual experience without attending to it or its content at all. Moreover when one does attend, it seems that the explanation of one's awareness of the relevant feature as *that* feature is, in part, that one is having an experience that represents it. But no such explanation is possible if the content of the experience is already conceptual.

If this is correct, then Albert's thoughts cannot possibly have the same content as the visual experiences of his sighted fellows. Albert's thoughts of color and shape are conceptual representations. They have conceptual content. Content-wise, they may *overlap* with the contents of visual experiences, since, as noted above, visual experience typically has some partly conceptual content. But there will be no identity. The nonconceptual content of visual experiences is content that cannot be captured in words. For where words enter, so do concepts.[7]

The Inverted Spectrum

Tom has a very peculiar visual system. His visual experiences are systematically inverted with respect to those of his fellows. When Tom looks at red objects, for example, what it is like for him is the same as what it is like for other people when they look at green objects and vice versa. This peculiarity is one of which neither he nor others are aware. Tom has learnt the meanings of colour words in the usual way and he applies these words correctly. Moreover his nonlinguistic behaviour is standard.

Now when Tom views a tomato, say, in good light his experience is phenomenally, subjectively different from the experiences you and I undergo. But his experience has the same representational content as ours. For his experience is the sort that is usually produced in him by viewing red objects and that usually leads him to believe that a red object is present. So he, like you and me, in viewing the tomato has an experience that represents the tomato as *red* (see Shoemaker 1975). The only way that Tom's experience can be subjectively different from yours and mine, then, is if it has a different visual Quale. This intrinsic phenomenal quality partly in virtue of which his experience represents the tomato as red cannot be the one partly in virtue of which our experiences represent it as red. Rather his is the one partly in virtue of which other experiences of ours represent grass and leaves, for example, as green.

One might respond to this argument by denying that a behaviourally undetectable inverted spectrum is possible.[8] There is another response available, however, that seems to me intuitively very satisfying. Contrary to what is claimed above, I believe that the difference between Tom and the rest of us when he views a tomato is that his experience, unlike ours, represents it as *green.* How is this possible? After all, the content of Tom's experience must be given to him, for the difference is a subjective one. But if the content is given to him then he must be introspectively aware that his experience represents the tomato as green. Unfortunately he is aware of no such thing. He sincerely asserts that the tomato is red and even that it looks red to him. Moreover, as was noted above, his experience is the sort that in him is typically produced by viewing red objects.

The answer, I maintain, is as follows: Introspection leads Tom astray. He forms a false belief about the content of his experience.[9] This content is certainly something *of* which he is introspectively aware but it is a content which he misclassifies. He takes it to be the content *red* and so he believes, on the basis of introspection, that he is undergoing an experience that represents red. In reality his experience represents green. *This* representational difference is what is responsible for the subjective difference between his experience and ours. Tom's mistake is due, of course, to the fact that he is unaware of his peculiarity. He does not know that his visual system is producing experiences with atypical contents; he thinks he is normal and he knows that the experience he undergoes viewing the tomato is subjectively like those he undergoes viewing other red objects. So he thinks that his experience represents red.

Perhaps it will be said that I haven't explained how Tom's experience can represent green when it is an experience of the subjective sort that is normally produced in him by viewing red objects and which normally produces in him the belief that something red is present. My reply is that Tom's experience certainly represents red, at the conceptual level. Tomatoes look to him *to be* red. They look to him *as if* they are red. Moreover, tomatoes look to Tom *like* other red things. But that is compatible with holding that, at the nonconceptual level, tomatoes look green to him. As Roderick Chisholm 1957 and Frank Jackson 1977 forcefully argued some years ago, a distinction needs to be drawn between locutions of the form '*X* looks *F* to *S*' and '*X* appears *F* to *S*,' where '*F*' expresses a sensory property, that is, a property of which one is directly aware via introspection, as one undergoes a visual experience, and the other 'looks' locutions noted above.

In my view, '*X* looks *F* to *S*,' given an appropriate '*F*,' is a paradigm of phenomenal talk, and it is best taken to express how *S*'s visual experience represents *X*, namely as having *F*. This is reflected in the intensionality of such talk in two ways. First, it can be true that *X* looks *F* to *S*, even if there is no *X*. Suppose, for example, that I bang my head and I see stars.[10] Here there are no stars, but it can still be true that the stars look bright to me. Secondly, it can be true that *X* looks *F* to *S* without *X*'s looking *G* to *S*, even if '*F*' and '*G*' are co-extensive. Suppose that, as it happens, everything purple is poisonous and everything poisonous is purple. Still, intuitively something, in looking purple to me, does not look poisonous. This example also illustrates the point about different 'looks' locutions. If I am aware of the connection between being purple and being poisonous, upon noting the apparent color of a given object before me, I may see it *as* poisonous; it may look to me *to be* poisonous. But it does not look poisonous, in the phenomenal sense of the term 'looks.'

Even though phenomenal 'looks' talk is intensional, in my view, it is not intensional to the same degree as propositional attitude contexts. And this is because, to repeat, basic visual experience is nonconceptual, as is experience generally. It seems plausible to suppose that for creatures like us, creatures with an evolutionary history, the phenomenal character of states like feeling pain or having a visual sensation of red is phylogenetically fixed. On this view, through learning we can change our beliefs, our thoughts, our judgements, but not (by and large) how things look and feel (in the phenomenal sense of these terms). Having acquired the concept microscope, say, we can come to see something as a microscope, but we do not need concepts simply to see. Once the receptor cells are matured, it suffices to open the eyes. No learning or training is involved. The phenomenal appearances are nonconceptual. Small children see pretty much what the rest of us see. Things look phenomenally to them pretty much as they do to adults, assuming no inverted spectra. They differ in *how* they see things, in what they see things *as*. They do not see that the kettle is boiling, the house as being dilapidated, the computer as malfunctioning.

In the case of Tom, given his visual abnormality, there are striking phenomenal differences from the rest of us. His visual system nonconceptually represents red things as green while conceptually representing them as red. The concept red Tom has is one he shares with you and me, notwithstanding these phenomenal differences (a concept that enables him to recognize red things and to discriminate them from things of other colors directly by sight on the basis of how they appear phenomenally).

Of course, this view entails that Tom is constantly misperceiving at the nonconceptual level, even though his color *judgements* are as accurate as the rest of us. And that, it may be urged, is impossible. But why? So-called 'normal misperceptions' occur with respect to shape, length, orientation. For example, in the Muller-Lyer illusion, two parallel lines of the same length look different lengths to normal perceivers in normal viewing conditions. Why not, then, in the case of Tom and color?

Suppose, however, Tom is not a lone invert. Color-qualia inversions are rife. Who now gets to undergo accurate color experiences? Who experiences ripe tomatoes as red, grass as green, and so on, at the nonconceptual level? It may seem that there is no nonarbitrary way of picking out a subpopulation of normal perceivers whose color experiences do not misrepresent. Any choice of a subpopulation may seem as good as any other. Unfortunately, if that is the case, then there is no fact of the matter about who is misrepresenting. So, it may seem, the example of rife phenomenal inversions shows that the attempt to do away with visual Qualia is misguided.

I grant that in the case of rife inversions, without further information, we all have an equal right to accurate color experiences. But this establishes nothing. Only if no further story is available under which some humans could end up being in the wrong while others remain in the right is there any problem here for the representationalist approach to visual qualia.

Such a story can be given in teleological terms.[11] Suppose, for example, there is a genetic defect in certain humans that are alive today, the result of which is that wires are crossed in their visual system, thereby inducing in them color experiences opposite to those that were present (in the same conditions) in most of their ancestors. Originally only a small subpopulation of the human species had the given defect, but now it has spread so that a sizeable number of us have it. These people have an experience of red when they see green things in daylight. They have an experience of green when they see red things in daylight, and so on. Their experiences are now tracking colors that are opposite on the hue circle to those tracked by their biologically normal ancestors. Since the visual systems of the members of this human subpopulation are not functioning as they were designed to do, the colors their sensory states *would* track, *were* they discharging their biological function, are not the colors they actually track. This is how misrepresentation arises. The nonconceptual visual states of these humans are not tracking the colors they were designed to track. So, error enters. Likewise, for other possible subpopulations. No obvious difficulty, then, for the rejection of visual Qualia.

Twin Earth

Jones is watching a cat. On Putnam's planet, Twin Earth, Jones' doppelganger is watching a creature that looks just like a cat but is genetically and biologically very different (see Putman 1975). Jones and Twin Jones are subject to retinal images that exactly match and their brains are in exactly the same physicochemical states. Intuitively, then, it may be urged, their visual sensations are phenomenally identical. But the contents of their experiences are different. Since Twin Jones has never seen or heard of cats (there aren't any cats on Twin Earth, only twin cats) and the beliefs he forms on the basis of his visual experiences are never of the type 'This is a cat,' Twin Jones' experience represents not that there is a cat but rather that there is a twin cat present. So the phenomenal sameness obtaining between Jones' and Twin Jones' visual experiences cannot be grounded in a sameness of content. Rather it must be grounded in the experiences sharing identical Qualia.

This argument forgets that Twin Jones' visual experience represents much more than just that a twin cat is present; it also represents the location of the twin cat relative to the viewer, its shape, colour, orientation, and a myriad of other surface details. These aspects of the content of Twin Jones' visual experience will also be found in the content of Jones' experience. I maintain that the phenomenal sameness obtaining between their visual experiences is traceable to these shared aspects.

Still, might not Jones and Twin-Jones differ with respect to their visual representation of color, even though they are microphysical duplicates? If inverted spectrum cases really are possible, then it certainly seems that Jones and Twin-Jones, given their different settings and (let us suppose) evolutionary histories, could differ with respect to their nonconceptual representation of color. So, if our account of the phenomenal is one which ties it to aspects of representational content, and nonconceptual content in particular, then we must reject the widely held view that subjective, phenomenal states of consciousness supervene on brain activity (as is implicitly supposed in the Twin earth argument above).

Why should that disturb us? Those who insist on such supervenience have no strong arguments for their view. The fact that Jones and Twin-Jones' phenomenal states are caused by the same brain states certainly does not show that their phenomenal states must be the same any more than does the fact that their beliefs are caused by the same brain states shows that their beliefs must be the same. Internal supervenience for the phenomenal is no more than a

dogma. And sleeping dogmas should not be left undisturbed. Its origin is the Cartesian view of experience as involving inner conscious ideas or pictures. Fix the neurophysiology and you fix the mental paint. Fix that and you fix the phenomenal. Not so, I claim. Phenomenology ain't in the head.

Peacocke's Puzzle Cases

In *Sense and Content* (Peacocke 1983), Christopher Peacocke presents a number of ingenious cases designed to show that sensory experiences have Qualia or, as he calls them, 'sensational properties.'[12] Peacocke's first case is as follows: Two trees of the same size are viewed, one twice as close as the other. Here, if the situation is normal, the visual experience represents the two trees as being of the same size. They look to the viewer the same size. But there is a sense in which the trees look different: the closer tree occupies a larger region in the visual field, and this, according to Peacocke, can only be accounted for nonrepresentationally via a sensational quality or Quale.

There is another possibility. The reason that the trees look different is, I believe, that the experience represents the nearer tree as having a facing surface that differs in its viewpoint-relative size from the facing surface of the further tree, even though it also represents the two trees as having the same viewpoint-independent size. The nearer tree (or its facing surface) is represented as being *larger from here*, while also being represented as being the same objective size as the further tree. There really are two different sorts of feature being represented, then, although they both are concerned with physical objects (or surfaces).

But what exactly is involved in one of two items being larger from here? The obvious answer is that the one item subtends a larger visual angle relative to the eyes of the viewer.[13]

Peacocke rejects this proposal on the grounds that experiences like mine can be had by people who lack the concept of a visual angle. My reply is that the perceptual experience represents the feature, being larger from here, nonconceptually. For a person to undergo an experience that represents one thing as larger relative to his viewing point than another, it suffices that the encoding feature of the experience (larger number of filled array cells, if the representational vehicle has an array-like structure) suitably track or causally covary with the instantiation of the viewpoint-relative relation. The person does not need to have any cognitive grasp of subtended angles.

The key claims I want to make, then, with respect to the tree case are these: (1) the nearer tree looks the same objective size as the further away tree while also looking larger from the given viewing position. (2) *X* looks *F* to *P* only if *P* undergoes a visual experience with respect to *X* that represents *F*. (3) Where the sense of 'looks' in (2) is phenomenal, the representation involved is nonconceptual. (4) The relevant nonconceptual, representational relation is a backward-looking tracking relation. Note that, on this account, the perceiver of the two trees is not the subject of any illusion or error: the nearer tree is just as it looks—both larger from here, the viewing position, and the same viewer-independent size as the further away tree.

Peacocke's second case appeals to a contrast between binocular and monocular vision. If I view a situation with both eyes and then close an eye, things will appear different to me. This difference, according to Peacocke, is not representational. Things are represented in just the same ways in both experiences. So, the difference must be due to a difference in Qualia.

The claim I reject here (not surprisingly) is the claim that there is no representational difference. When I view the situation with both eyes, I see a little more at the periphery of my visual field and there is an increase in how determinately my experience represents object depth. An appeal to Qualia is not required.

Peacocke's third example is a case in which a wire cube is seen first as having one face in front of the other and then with the relative positions of the two faces reversed (see Figure 27.1). Although there is a change in the experience here, something in the experience remains the

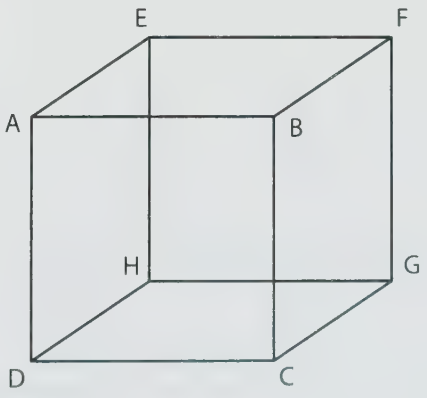


Figure 27.1

same. This constant feature of the experience is, Peacocke maintains, a sensational quality.

The obvious response to this example is to concede the point that something in the experience remains the same but to explain this fact representationally by holding that both before and after the 'aspect' switch, the experience represents the cube as having various unchanging spatial properties relative to the given point of view. For example, both before and after the switch, side ABCD is represented as being lower than and somewhat to the left of side EFGH, side AEHD is represented as being level with and wholly to the left of side BFGC, and so on.

Other aspect switches are no more problematic for my position, I might add. Consider, for example, the pattern in Figure 27.2 (which Peacocke mentions a little later). We may see this pattern either with the dots running from the bottom to the top or from the left to the right. How is this to be accounted for? Answer: The pattern looks composed of columns of dots in the one case and it looks composed of rows of dots in the other. In the former case, the experience represents the pattern as composed of columns; in the latter, rows. The experiences are phenomenally different, then, because they represent different groups of dots. This representational difference explains why the perceiver will judge the pattern similar to Figure 27.3(a) in the former instance and similar to Figure 27.3(b) in the latter. The overall conclusion I reach, then, is that there is no need to postulate visual QUALIA in order to account for the subjective aspects of our visual experiences.

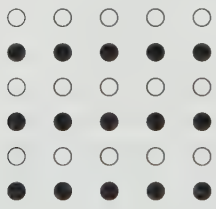


Figure 27.2

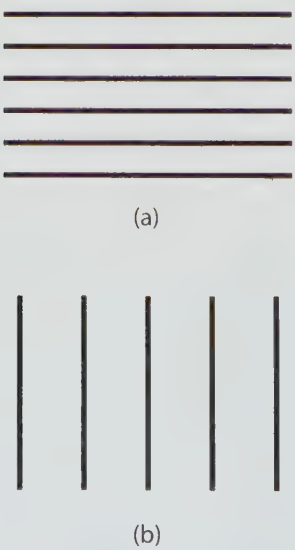


Figure 27.3

III

Consider again the hypothesis that necessarily visual experiences with the same representational contents have the same phenomenal character or 'feel.' If this hypothesis is true, as I am claiming, it seems implausible to suppose that its truth is just a brute fact. The natural explanation is that phenomenal 'feel' is itself representational content of a certain sort, representational content that meets certain further conditions.

In Tye 1995 and 2000, I argue that phenomenal content, as we might call it, is PANIC: poised, abstract, nonconceptual, intentional (or representational) content. The requirement that phenomenal content be nonconceptual is compatible with the well established thesis that the way in which one conceives of a scene may causally influence the phenomenal character of one's visual experience of it. In such cases, there is always a difference in the features nonconceptually represented by the experience.[14] So, it is not necessary, to concede that concepts are integral to the 'feel' of the experience. As far as the 'feel' goes, the relevant represented features are not ones for which the subject need possess corresponding concepts at all.

In saying that phenomenal content must be *abstract*, I mean that it must be content into which no particular concrete objects or surfaces enter. This is required by the case of hallucinatory experiences, for which no concrete objects need be present at all; and it is also demanded by cases in which different objects look exactly alike phenomenally. What is crucial to phenomenal character is, I claim, the representation of general features or properties.

The requirement that phenomenal content be (suitably) *poised* is essentially a functional role one. The key idea is that experiences, qua bearers of phenomenal character, play a certain

distinctive functional role. They arise at the interface of the nonconceptual and conceptual domains, and they stand ready and available to make a direct impact on beliefs and/or desires. For example, how things phenomenally look typically cause beliefs as to how they are, if attention is properly focused. States with nonconceptual content that are not so poised lack phenomenal character. Consider, for example, states generated in vision that nonconceptually represent changes in light intensity. These states are not appropriately poised. They arise too early, as it were, in the information processing. The information they carry is not directly accessible to the relevant cognitive centers.[15]

Why does it matter whether visual Qualia can be avoided? One answer, I suggest, is that with the rejection of visual Qualia certain aspects of visual experience become less puzzling. Let me explain.

Any adequate account of the subjective or phenomenal aspects of our visual states ought to yield an understanding of why those states have those aspects. Why, for example, does having a visual experience of blue 'feel' the way it does and not some other way? It is hard to see how any satisfying answer can be given to this question if the phenomenal aspects of such experiences derive from visual Qualia.

Suppose, for example, that there are visual Qualia and that such Qualia are nonphysical and irreducible. Then the 'felt' aspect of the visual experience of blue is a matter of its having a special, nonphysical property. It is the presence of this property that gives the visual experience its distinctive 'feel.' Does this really offer us any enlightenment? Apart from the usual concerns about the emergence and causal role of such properties we may still wonder why the visual experience that has the content blue is associated with this irreducible felt quality rather than some other—why, for example, it does not have the felt quality of experiences that represent red. This surely is an impenetrable mystery.

Suppose now that visual Qualia are physically reducible. Then the 'felt' aspect of the visual sensation of blue is a matter of its having a certain physio-chemical property. That is, I think, an improvement on the above alternative—it dissolves the worry about the causal role of Qualia, for example—but again it does not begin to explain why the visual experience that represents blue should 'feel' as it does.

On the proposal I have made there is a simple explanation. Introspection tells us that the visual experience that represents blue differs phenomenally from the visual experience that represents red. This 'felt' difference is, I claim, solely a matter of content. Since the colours represented by the two experiences are different, the experiences themselves are introspectively distinguishable. The reason, then, that the visual experience of blue 'feels' as it does is that it could not 'feel' any other way. The 'felt' aspect simply cannot be divorced from the representational aspect.

The onus now lies with the advocate of Qualia. I have tried to show that the rejection of visual Qualia is defensible against a variety of objections and that it is not only intuitively satisfying but also well motivated.[16]

REFERENCES

Block, N., "The Photographic Fallacy in the Debate about Mental Imagery," *Nôus* 83 (1983): pp. 654–64.

Chisholm, R. *Perceiving: A Philosophical Study* (Ithaca, NY: Cornell University Press, 1957).

DeBellis, M., "The Representational Content of Musical Experience," *Philosophy and Phenomenological Research* 51 (1991): pp. 303–24.

Dretske, F. *Naturalizing the Mind* (Cambridge, MA: Bradford Books/MIT Press, 1995).

Harman, G., "The Intrinsic Quality of Experience," in *Philosophical Perspectives*, 4:31-52, J. Tomberlin, ed., (Northridge, CA: Ridgeview Publishing Co. 1990).

Jackson, F. *Perception* (Cambridge, UK: Cambridge University Press, 1977).

Lycan, W., "Layered Perceptual Representation," in *Philosophical Issues*, 7, E. Villeneuva (ed), (Northridge, CA: Ridgeview Publishing Co., 1996).

McDowell, J., "The Content of Perceptual Experience," *Philosophical Quarterly* 55(175): 206-13 (April 1994).

Moore, G., "The Refutation of Idealism," in *Philosophical Studies* (London: Routledge & Kegan Paul, 1922).

Peacocke, C. *Sense and Content* (Oxford: Oxford University Press, 1983).

_____ "Scenarios, Concepts, and Perception," in *The Contents of Experience: Essays on Perception*, T. Crane, ed., (Cambridge, UK: Cambridge University Press, 1992), pp. 105–35.

Putnam, H., "The Meaning of 'Meaning,'" in *Language, Mind and Knowledge: Minnesota Studies in the Philosophy of Science*, vol. VII, Keith Gunderson, ed., (Minneapolis: University of Minnesota Press, 1975).

Shoemaker, S., "Functionalism and Qualia," *Philosophical Studies* 27 (1975): pp. 291–315.

Tye, M., "The Picture Theory of Mental Images," *Philosophical Review* 88 (1988): pp. 497–520.

_____ *The Imagery Debate* (Cambridge, MA: Bradford Books/MIT Press, 1991).

_____ *Ten Problems of Consciousness* (Cambridge, MA: Bradford Books/MIT Press, 1995).

_____ "Perceptual Experience Is a Many-Layered Thing," *Philosophical Issues*, 7, E. Villeneuva (ed), (Northridge, CA: Ridgeview Publishing Co., 1996).

_____ *Consciousness, Color, and Content* (Cambridge, MA: Bradford Books/MIT Press, 2000).

_____ "Representationalism and the Transparency of Experience," *Nôus* (2002).

NOTES

1. If they represent anything at all. Some defenders of Qualia deny that after-image experiences are representational. See Section II.
2. For more here, see Section III.
3. In something like the manner suggested by Stephen Kosslyn for mental images. For a summary of Kosslyn's views here, see Tye 1988 and Tye 1991.
4. Attaching to an image that is the mental object of the experience.
5. A case like that of Albert was suggested to me in conversation by Stephen Stich.
6. See here Peacocke 1992 for some plausible examples.
7. In giving this reply, I am not supposing that a state whose nonconceptual content duplicates the nonconceptual content of a given visual experience *v* is thereby a state whose phenomenology duplicates that of *v* or even that it is a state for which there is anything it is like to undergo it at all. In my view, a necessary condition of a state's content endowing it with phenomenal character is that the content be nonconceptual. Having nonconceptual content is not sufficient, however. The content must also be abstract and suitably poised. See here Section III.
8. This is the line taken by Gilbert Harman 1990. One problem that confronts such a line is that even if Tom's peculiarity is ultimately behaviourally detectable, it appears that some possible inversions are not, e.g., inversions pertaining to the experiences of creatures who see the world in black, white, and varying shades of grey. See here Shoemaker 1975. For a reply, see Tye 1995.
9. This position, together with an internalist conception of knowledge which requires Tom to cite the belief that his experience represents a red object in any adequate justification of the claim that the tomato before him is red, entails that he does not know that the tomato is red. Indeed, more generally, it entails that he does not know the colour of anything on the basis of vision despite his excellent performance. (I owe this point to Sydney Shoemaker.) Since the conclusion reached here is obviously false, I maintain that the above internalist conception of knowledge must be rejected.
10. This is the phenomenological use of the term 'see,' not the success use. See Tye 2000, chapter 4.
11. I endorse such a view only for creatures with an evolutionary history. See here Tye 2000. For an unqualified endorsement of the view, see Dretske 1995.
12. The replies I give below to Peacocke's examples are influenced by DeBellis 1991 and Harman 1990.
13. For more here, see Tye 1996. For an alternative reply, see Lycan 1996. This reply is criticized in Tye 1996.
14. For more here, see Tye 1995, pp. 140–41; Tye 2000.
15. Inclusion of the 'poised' condition in the account of phenomenal character entails that people with the real-world psychological impairment of blindsight (not to be confused with Albert's condition) do not have visual states with phenomenal character with respect to the blind portions of their visual fields.
16. 1992 version: I am grateful especially for comments by Chris Peacocke and Sydney Shoemaker. Revised version (2002): The current essay differs substantially in a number of places from the original one. In this connection, I am indebted to David Chalmers for some helpful suggestions. The revisions bring the paper more in line with my current view (Tye 2000), but in an attempt to increase accessibility for the present collection, many subtleties and qualifications have been passed over. For a detailed and careful statement of the appeal to the transparency or diaphanousness of experience on behalf of representationalism, see Tye 2002.

Illusionism as a Theory of Consciousness

Keith Frankish

> So, if he's doing it by divine means, I can only tell him this: 'Mr Geller, you're doing it the hard way.' (James Randi, 1997, p. 174)

Theories of consciousness typically address the hard problem. They accept that phenomenal consciousness is real and aim to explain how it comes to exist. There is, however, another approach, which holds that phenomenal consciousness is an illusion and aims to explain why it *seems* to exist. We might call this *eliminativism* about phenomenal consciousness. The term is not ideal, however, suggesting as it does that belief in phenomenal consciousness is simply a theoretical error, that rejection of phenomenal realism is part of a wider rejection of folk psychology, and that there is no role at all for talk of phenomenal properties—claims that are not essential to the approach. Another label is 'irrealism,' but that too has unwanted connotations; illusions themselves are real and may have considerable power. I propose 'illusionism' as a more accurate and inclusive name, and I shall refer to the problem of explaining why experiences seem to have phenomenal properties as the *illusion problem*.[1]

Although it has powerful defenders—preeminently Daniel Dennett—illusionism remains a minority position, and it is often dismissed out of hand as failing to 'take consciousness seriously' (Chalmers, 1996). The aim of this article is to present the case for illusionism. It will not propose a detailed illusionist theory, but will seek to persuade the reader that the illusionist research programme is worth pursuing and that illusionists do take consciousness seriously—in some ways, more seriously than realists do.[2]

1. Introducing Illusionism

This section introduces illusionism, conceived as a broad theoretical approach which might be developed in a variety of ways.

1.1. Three Approaches to Phenomenal Consciousness

Suppose we encounter something that seems anomalous, in the sense of being radically inexplicable within our established scientific world-view. Psychokinesis is an example. We would have, broadly speaking, three options. First, we could accept that the phenomenon is real and explore the implications of its existence, proposing major revisions or extensions to our science, perhaps amounting to a paradigm shift. In the case of psychokinesis, we might posit previously unknown psychic forces and embark on a major revision of physics to accommodate them. Second, we could argue that, although the phenomenon is real, it is not in fact anomalous and can be explained within current science. Thus, we would accept that people really can move things with their unaided minds but argue that this ability depends on known forces, such as electromagnetism. Third, we could argue that the phenomenon is illusory and set about investigating how the illusion is produced. Thus, we might argue that people who seem to have psychokinetic powers are employing some trick to make it seem as if they are mentally influencing objects.

The first two options are *realist* ones: we accept that there is a real phenomenon of the kind there appears to be and seek to explain it. Theorizing may involve some modest reconceptualization of the phenomenon, but the aim is to provide a theory that broadly vindicates our pre-theoretical conception of it. The third position is an *illusionist* one: we deny that the phenomenon is real and focus on explaining the appearance of it. The options also differ in explanatory strategy. The first is *radical*, involving major theoretical revision and innovation, whereas the second and third are *conservative*, involving only the application of existing theoretical resources.

Turn now to consciousness. Conscious experience has a subjective aspect; we say it is *like something* to see colours, hear sounds, smell odours, and so on. Such talk is widely construed to mean that conscious experiences have introspectable qualitative properties, or 'feels,' which determine what it is like to undergo them. Various terms are used for these putative properties. I shall use 'phenomenal properties,' and, for variation, 'phenomenal feels' and 'phenomenal character,' and I shall say that experiences with such properties are *phenomenally conscious*. (I shall use the term

From *Journal of Consciousness Studies* 23: 11–39, 2016.

'experience' itself in a functional sense, for the mental states that are the direct output of sensory systems. In this sense it is not definitional that experiences are phenomenally conscious.) Now, phenomenal properties seem anomalous. They are sometimes characterized as simple, ineffable, intrinsic, private, and immediately apprehended, and many theorists argue that they are distinct from all physical properties, inaccessible to third-person science, and inexplicable in physical terms. (I use 'physical' in a broad sense for properties that are either identical with or realized in microphysical properties.) Again, there are three broad options.

First, there is radical realism, which treats phenomenal consciousness as real and inexplicable without radical theoretical innovation. In this camp I group dualists, neutral monists, mysterians, and those who appeal to new physics. Radical realists typically stress the anomalousness of phenomenal properties, their resistance to functional analysis, and the contingency of their connection to their neural correlates. Second, there is conservative realism, which accepts the reality of phenomenal consciousness but seeks to explain it in physical terms, using the resources of contemporary cognitive science or modest extensions of it. Most physicalist theories fall within this camp, including the various forms of representational theory. Both radical and conservative realists accept that there is something real and genuinely qualitative picked out by talk of the phenomenal properties of experience, and they adopt this as their explanandum. That is, both address the hard problem.[3]

The third option is illusionism. This shares radical realism's emphasis on the anomalousness of phenomenal consciousness and conservative realism's rejection of radical theoretical innovation. It reconciles these commitments by treating phenomenal properties as illusory. Illusionists deny that experiences have phenomenal properties and focus on explaining why they seem to have them. They typically allow that we are introspectively aware of our sensory states but argue that this awareness is partial and distorted, leading us to misrepresent the states as having phenomenal properties. Of course, it is essential to this approach that the posited introspective representations are not themselves phenomenally conscious ones. It would be self-defeating to explain illusory phenomenal properties of experience in terms of real phenomenal properties of introspective states. Illusionists may hold that introspection issues directly in dispositions to make phenomenal judgments—judgments about the phenomenal character of particular experiences and about phenomenal consciousness in general. Or they may hold that introspection generates intermediate representations of sensory states, perhaps of a quasi-perceptual kind, which ground our phenomenal judgments. Whatever the details, they must explain the content of the relevant states in broadly functional terms, and the challenge is to provide an account that explains how real and vivid phenomenal consciousness seems. This is the illusion problem.

1.2. Illusionism Strong and Weak

Illusionism makes a very strong claim: it claims that phenomenal consciousness is illusory; experiences do not really have qualitative, 'what-it's-like' properties, whether physical or non-physical. This should be distinguished from a weaker view according to which some of the supposed *features* of phenomenal consciousness are illusory. Many conservative realists argue that phenomenal properties, though real, do not possess the problematic features sometimes ascribed to them, such as being ineffable, intrinsic, private, and infallibly known. Phenomenal feels, they argue, are physical properties which introspection misrepresents as ineffable, intrinsic, and so on. We might call this *weak illusionism*, in contrast to the strong form advocated here. (It might equally be called *weak realism*.)[4]

On the face of it, weak and strong illusionism are similar. Both hold that experiences have distinctive physical properties that are misrepresented by introspection. There is a crucial difference, however. Weak illusionism holds that these properties are, in some sense, genuinely *qualitative*: there really are phenomenal properties, though it is an illusion to think they are ineffable, intrinsic, and so on. Strong illusionism, by contrast, denies that the properties to which introspection is sensitive are qualitative: it is an illusion to think there are phenomenal properties at all.

We can highlight the difference by introducing the notion of a *quasi-phenomenal property*. A quasi-phenomenal property is a non-phenomenal, physical property (perhaps a complex, gerrymandered one) that introspection typically misrepresents as phenomenal. For example, quasi-phenomenal redness is the physical property that typically triggers introspective representations of phenomenal

redness.[5] There is nothing phenomenal about such properties—nothing 'feely' or qualitative—and they present no special explanatory problem. Strong illusionists hold that the introspectable properties of experience are merely quasi-phenomenal ones. But weak illusionists cannot agree. If experiences have only quasi-phenomenal properties, then it would be misleading to say that phenomenal properties are real, just as it would be misleading to say that psychokinetic powers are real if all people can do is create the illusion of having them.

The moral is that if weak illusionism is not to collapse into strong illusionism, then it must employ a concept of phenomenality stronger than that of quasi-phenomenality. Indeed, one motive for advancing the strong illusionist position is to force conservative realists to face up to the challenge of articulating a concept of the phenomenal that is both stronger than that of quasi-phenomenality and weak enough to yield to conservative treatment. I doubt this is possible (see Frankish, 2012) and, if it is not, then radical realism and strong illusionism will be the only options. In what follows, 'illusionism' will always mean *strong illusionism*.

1.3. Some Analogies

Illusionists offer various analogies to illustrate their view. Dennett compares consciousness to the user illusions created by the graphical interfaces through which we control our computers (Dennett, 1991, pp. 216–20, 309–14). The icons, pointers, files, and locations displayed on a computer screen correspond in only an abstract, metaphorical way to structures within the machine, but by manipulating them in intuitive ways we can control the machine effectively, without any deeper understanding of its workings. The items that populate our introspective world have a similar status, Dennett suggests. They are metaphorical representations of real neural events, which facilitate certain kinds of mental self-manipulation but yield no deep insight into the processes involved. (Dennett stresses the limits of the interface analogy. There is no internal display for the benefit of a conscious user; the illusion is a product of the limited access relations between multiple nonconscious subsystems and it manifests itself in our personal-level intuitions and judgments about our inner lives.)

Rey cites cases where stabilities in our reactions to the world induce us to project corresponding properties onto the world (Rey, 1995, pp. 137–9). For example, our stable personal concerns and reactions to others lead us to posit stable, persisting selves as their objects. Similarly, Rey suggests, our representations of our own and others' experiences lead us to posit simple mental phenomena corresponding to them. Take pain, for example. We have a 'weak,' functional concept of pain, which includes links both to sensory representations of pain encoding information about intensity, apparent location, and so on, and to third-person representations of pain behaviour in others. Reflecting on our own and others' pains, we then develop a 'strong,' qualitative concept of pain as the thing that is the immediate object of our pain experiences and the cause of pain behaviour in others.

Humphrey compares sensations to impossible objects, such as the Penrose triangle, depicted on the left of Figure 28.1. Such an object cannot exist in three-dimensional space, but the illusion of it can be created by the object on the right, which Humphrey calls the *Gregundrum*, after its creator Richard Gregory. From most perspectives the Gregundrum appears an ungainly construction, but from just the right angle it looks like a solid Penrose triangle. Consciousness, Humphrey proposes, involves

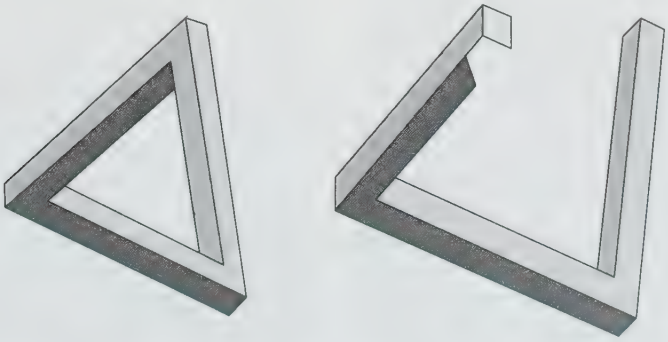


Figure 28.1 The Penrose triangle and the Gregundrum.

an analogous illusion. Our brains create an 'ipsundrum'—a neural state that appears relatively unremarkable from other perspectives but generates the illusion of phenomenality when viewed introspectively (Humphrey, 2011, chapter 2). Phenomenal consciousness is a '*fiction of the impossible*' (*ibid.*, p. 204)—a magic trick played by the brain on itself. (Talk of illusion should not be taken to indicate a defect in introspection; Humphrey argues that the illusion is highly adaptive; Humphrey 2006; 2011.)

Pereboom draws a comparison with secondary qualities, such as colours (Pereboom, 2011, pp. 15–40). It is arguable that sensory perception represents colours as properties of external objects resembling the sensations they produce in us. Since objects lack such properties, sensory perception universally misrepresents objects in this respect. Similarly, Pereboom suggests, introspection may universally misrepresent phenomenal properties as having qualitative natures they do not in fact have. (By 'phenomenal properties' he means the distinctive introspectable properties of conscious experiences, whatever they may be. Phenomenal properties in this sense may be merely quasi-phenomenal.) Pereboom calls this the *qualitative inaccuracy hypothesis*, and he argues that it is an open possibility. If it seems less credible than the parallel hypothesis about secondary qualities, Pereboom suggests, this is because we cannot check the accuracy of introspection, as we can that of perception, by adopting different vantage points, using measuring instruments, and so forth (ibid., p. 23).

These analogies all illustrate the basic illusionist claim that introspection delivers a partial, distorted of view of our experiences, misrepresenting complex physical features as simple phenomenal ones. Sensory states have complex chemical and biological properties, representational content, and cognitive, motivational, and emotional effects. We can introspectively recognize these states when they occur in us, but introspection doesn't represent all their detail. Rather, it bundles it all together, representing it as a simple, intrinsic phenomenal feel. Applying the magic metaphor, we might say that introspection sees the complex sleight-of-hand performed by our sensory systems as a simple magical *effect*. And, as with a conjuring trick, the illusion depends on what the audience does not see as much as what they do. In another analogy, Rey compares our introspective lives to the experience of a child in a dark cinema who takes the cartoon creatures on screen to be real (Rey 1992, p. 308). The illusion depends on what the child doesn't see—on the fact that their visual system does not register individual frames as distinct images. Cinema is an artefact of the limitations of vision, and, illusionists may say, phenomenal consciousness is an artefact of the limitations of introspection.

The analogy with visual illusions also holds with respect to cognitive penetrability. Forming the theoretical belief that phenomenal properties are illusory does not change one's introspective representations, and one remains strongly disposed to make all usual phenomenal judgments (and perhaps does still make them at some level). As with perceptual illusions, this may indicate that the phenomenal illusion is an adaptive one, which has been hardwired into our psychology. (However, it may be possible to dispel the illusion partially through indirect means, such as meditation and hypnotic suggestion; see, for example, Blackmore 2011.)

The analogies also indicate some dimensions along which illusionist theories may differ. One concerns the sensory states that are the basis for the illusion. On most accounts, I assume, these will be representational states, probably modality-specific analogue representations encoding features of the stimulus, such as position in an abstract quality space, egocentric location, and intensity. Accounts will differ, however, on the details of their content, functional role, relation to attentional processes, and so on. Theories will also differ as to which properties of these states are responsible for the illusion of phenomenality (their quasi-phenomenal properties). Is introspection sensitive only to the content of sensory states, or are we also aware of properties of their neural vehicles? Do the reactions and associations evoked by our sensory states also contribute to the illusion of phenomenality?

Relatedly, there are questions about our introspective access to our sensory states. Do we have internal monitoring mechanisms that generate representations of sensory states, and if so what sort of representations do they produce? (Are they thoughts about sensory states or perceptions of their neural vehicles?) Are the introspective representations conscious or unconscious? (They are not phenomenally conscious, of course, but they could be conscious in the psychological sense of being globally available.) Are sensory states continually monitored or merely available to monitoring? Is the introspectability of sensory states a matter of internal access and influence rather than internal

monitoring?[6] There are many options here and parallels with higher-order representational theories of consciousness, some of which might be reformulated as illusionist ones.

1.4. Outward-Looking Illusionism?

I characterized illusionism as the view that phenomenal consciousness is an *introspective* illusion, reflecting the widely held view that phenomenal properties are properties of experience. This may be too restrictive, however. Some theorists hold that experience is *transparent*: when we attend to our experiences, we are aware only of properties of their objects. Thus, redness is experienced as a property of surfaces, pain as a property of parts of our bodies, and so on (e.g. Harman 1990; Tye 1995, 2000). This points to the possibility of an outward-looking illusionism, on which experience misrepresents distal stimuli as having phenomenal properties. Vision, for example, would represent objects as having illusory phenomenal colours as well as real physical colours (for a view of this kind, see Hall 2007).

This view can be regarded as a variant of standard, inward-looking illusionism, differing principally on where the illusory phenomenal properties are represented as being located. And, like the inward-looking version, it may posit processes of internal monitoring. The illusion of phenomenality may involve a combination of introspection and projection, in which we both misrepresent features of experience as phenomenal and then re-represent these illusory properties as properties of the external world, mistaking complex physical properties of our sensory states for simple phenomenal properties of external objects (Humphrey 2011, chapter 7). In what follows, I shall focus on the inward-looking form of illusionism, though most points will apply to both.

1.5. Illusionism and Grand Illusion

Illusionism should be distinguished from the thesis that the visual world is a grand illusion (Noë 2002). The latter holds that conscious visual experience is far less stable and detailed than we suppose, as is revealed by experiment and careful introspection. Illusionism, by contrast, is a thesis about conscious experience generally and concerns its nature, not its extent. One could hold that the visual world *is* stable and detailed while still claiming that it involves an illusion in the sense discussed here.

Nevertheless, evidence for the grand illusion view, such as the existence of change blindness, does lend support to illusionism. If we regularly overestimate the extent and stability of our conscious visual experience, then it is possible to be under a kind of illusion about one's own phenomenal consciousness. Moreover, as Dennett shows, phenomena such as change blindness undermine familiar intuitions about phenomenal properties, suggesting that our conception of them is incoherent and the properties themselves consequently illusory (e.g. Dennett 2005, pp. 82–91).

1.6. Illusionism and Eliminativism

Does illusionism entail eliminativism about consciousness? Is the illusionist claiming that we are mistaken in thinking we have conscious experiences? It depends on what we mean by 'conscious experiences.' If we mean experiences with phenomenal properties, then illusionists do indeed deny that such things exist. But if we mean experiences of the kind that philosophers *characterize* as having phenomenal properties, then illusionists do not deny their existence. They simply offer a different account of their nature, characterizing them as having merely quasi-phenomenal properties. Similarly, illusionists deny the existence of phenomenal consciousness properly so-called, but do not deny the existence of a form of consciousness (perhaps distinct from other kinds, such as access consciousness) which consists in the possession of states with quasi-phenomenal properties and is commonly mischaracterized as phenomenal. Henceforth, I shall use 'consciousness' and 'conscious experience' without qualification in an inclusive sense to refer to states that might turn out to be either genuinely phenomenal or only quasi-phenomenal. In this sense realists and illusionists agree that consciousness exists.

Do illusionists then recommend eliminating talk of phenomenal properties and phenomenal consciousness? Not necessarily. We might reconceptualize phenomenal properties as quasi-phenomenal ones. Recall Pereboom's analogy with secondary qualities. The discovery that colours are mind-dependent did not lead scientists to deny that objects are coloured. Rather, they reconceptualized colours as the properties that cause our colour sensations. Similarly, we might respond to the discovery that experiences lack phenomenal properties by reconceptualizing phenomenal properties as the properties

that cause our representations of phenomenal feels—that is, quasiphenomenal properties.[7] This could invite confusion, however, given how tightly the notion of phenomenality is bound up with dualist intuitions, and in scientific work it might be wiser to abandon talk of phenomenal properties and phenomenal consciousness altogether.

In everyday life, however, we would surely continue to talk of the feel or quality of experience in the traditional, substantive sense. As subjects of experience, our interest is in how things seem to us introspectively—the illusion itself, not the mechanisms that cause it. Such talk may fail to pick out real properties, but it is not empty or pointless. Consider another analogy. Having watched a performance of *King Lear*, Lucy remarks, 'Lear's anguish in the final scenes was heart-breaking.' What is she talking about? There was (we may suppose) no anguish on stage at all, only the artful illusion of it. And it would be implausible to construe Lucy as referring to the *cause* of this illusion—the actor's words and gestures (quasi-anguish, as it were). The words and gestures were not themselves heart-breaking. The answer, of course, is that Lucy is referring to a fictional agony, entering into the world of the play and responding to the emotions of the characters as if they were real. (And in doing so, we might add, she is not making an error but appreciating the very point of the performance.) Everyday talk about the quality of experience should, I suggest, be construed similarly. Of course, most people do not regard their phenomenology as illusory; they are like naïve theatregoers who take the action on stage for real. But if illusionists are right, then cognitive scientists should treat phenomenological reports as fictions—albeit ones that provide clues as to what is actually occurring in the brain.[8]

1.7. Zombies and What It Is Like

Are illusionists claiming that we are (phenomenal) zombies? If the only thing zombies lack is phenomenal consciousness properly so called, then illusionists must say that, in this technical sense, we are zombies. However, zombies are presented as creatures very different from ourselves—ones with no inner life, whose experience is completely blindsighted. As Chalmers puts it, 'There is nothing it is like to be a zombie. . . all is dark inside' (Chalmers 1996, pp. 95–6). And illusionists will not agree that this is a good description of us. Rather, they will deny the equivalence between having an inner life and having phenomenal consciousness. Having the kind of inner life we have, they will say, consists in having a form of introspective self-awareness that creates the illusion of a rich phenomenology.

But aren't phenomenal properties precisely what makes experience *like something*? That is certainly a common way of construing what-it's-like talk, but there is another way. Illusionists can say that one's experiences are like something if one is aware of them in a functional sense, courtesy of introspective representational mechanisms. Indeed, this is a plausible reading of the phrase; experiences are like something for a creature, just as external objects are like something for it, if it mentally represents them to itself. Illusionists agree that experiences are like something in this sense, though they add that the representations are non-veridical, misrepresenting experiences as having phenomenal properties (what-it's-likeness in the first sense). And in this second sense there *is* something it is like to be a zombie, since zombies have introspective mechanisms functionally identical to our own. When we imagine zombies as being different from us, we are—illegitimately—imagining creatures with different introspective capacities.

It may be objected that we can imagine a creature representing itself as having phenomenal properties while still lacking an inner life. Zombies believe they are phenomenally conscious (in some sense at least; arguably, they lack full-blown phenomenal concepts; Chalmers 1996, 2003). But—it may be said—this does not give them an inner life like ours. I am not sure this is obvious. Consider the grand illusion view again. Our sense that our visual field is uniformly rich and detailed may be a sort of cognitive illusion, reflecting expectations and assumptions about the information that vision provides, and our sense of having a rich phenomenology might be a similar cognitive illusion. But in any case the illusionist need not claim that the illusion depends solely on the possession of certain propositional attitudes. Rather, they may say, it depends on a complex array of introspectable sensory states, which trigger a host of cognitive, motivational, and affective reactions. If we knew everything about these states, their effects, and our introspective access to them, then, illusionists say, we could not clearly imagine a creature possessing them without having an inner life like ours.

Of course, it is easy to say that. Illusionists need to explain how it can be true. That is, they need to solve the illusion problem. But it would be begging the question against illusionism to assume that it cannot be done.

2. Motivating Illusionism

This section motivates illusionism, sketching its advantages over radical realism and conservative realism and then adding some positive arguments in its favour. It does not aim to present a watertight case for illusionism but simply to show that the view has strong attractions.

2.1. Against Radical Realism

I take it there is a presumption in favour of conservatism in science: we should not make radical theoretical moves if modest ones will do. Of course, when it comes to consciousness many are confident that modest moves *won't* do, but that is what conservative theorists deny. The principle of conservatism should apply with special force, I suggest, when the pressure for radical innovation comes from a parochial, anthropocentric source, such as introspection. Introspection delivers a view of ourselves that is peculiarly vivid and compelling and that seems radically at odds with that of the physical sciences. It *might* give us access to an aspect of reality inaccessible to third-person science. (Though even if it did, it is hard to see how we could develop a science of that aspect.[9]) But it might merely give us an unusual perspective on the same reality—a perspective that is partial and distorted and deceives us into thinking that our experiences are resistant to conservative explanation.

In addition, a conservative approach is much better placed to account for the *psychological significance* of consciousness. By the psychological significance of a mental event, I mean its cumulative cognitive, motivational, emotional, and other psychological effects across various contexts. The common-sense view is that the way our experiences feel has huge psychological significance. Sensations entice us, guide us, move us, warn us, and the memory and anticipation of them are powerful motivators. Not only this, they hugely *enrich* life. As Humphrey stresses, we relish sensation for its own sake, and this relish shapes our behaviour in profound ways (Humphrey 2011). But this assumes that experiences affect us in virtue of how they feel. And it is hard for radical theorists to vindicate this assumption. Non-physical properties can have no effects in a world that is closed under causation, as ours appears to be, and the mind sciences show no independent need to refer to exotic physical processes, such as quantum-mechanical ones. The threat of epiphenomenalism hangs over radical theories. Some radical theorists respond by arguing that phenomenal properties are intrinsic to basic physical entities and thus intimately involved in physical causal processes (e.g. Strawson 2006). However, even if this proposal does dispel the threat (which is doubtful; Howell 2015), it involves huge profligacy with phenomenal properties and preserves the potency of consciousness only at the cost of making all physical causation phenomenal.

2.2. Against Conservative Realism

Conservative realism promises to capture the common-sense view of consciousness, accepting the reality of phenomenal properties but identifying them with causally potent, physical properties. However, it is an unstable position, continually on the verge of collapsing into illusionism.

The central problem, of course, is that phenomenal properties seem too weird to yield to physical explanation. They resist functional analysis and float free of whatever physical mechanisms are posited to explain them. (In practice this becomes almost definitional of phenomenal consciousness; any physicalist theory can be rejected as missing out the essential qualitative element.) The arguments are well-known, and I shall not repeat them here.

Many physicalists respond by arguing that our anti-physicalist intuitions arise from the way we conceptualize phenomenal properties in introspection—a tactic known as the *phenomenal concept strategy* (e.g. Hill 1997; Hill and McLaughlin 1999; Loar 1990; Papineau 2002). The idea is that phenomenal concepts have an especially intimate link to their referents and lack *a priori* connections to physical concepts. (They are typically characterized as either demonstrative, recognitional, or quotational.) This intimacy and isolation, it is argued, give rise to anti-physicalist intuitions, even though phenomenal properties are physical ones. It is doubtful, however, that this really relieves the pressure on conservative realism. For the concepts must still be *phenomenal* ones (Tartaglia 2013, p. 828). If they are recognitional concepts, for example, they must be recognitional

concepts for the *feel* of experiences. The concept of a mere introspectable *something*, which might or might not be qualitative, is not a genuine phenomenal concept, and if we conceptualized the properties of experience in that way, we would not feel any resistance to thinking of them as physical (a bare something might as easily be physical as non-physical). But if phenomenal concepts refer to feels, then the challenge to conservative realists remains. They must either explain how these feels can be physical or accept that phenomenal concepts misrepresent experience, as illusionists claim.

Looking at proposed reductive explanations themselves, the pressures towards illusionism become even clearer. As noted earlier, most physicalists adopt a weakly illusionist view, denying that phenomenal properties are private, intrinsic, and ineffable and employing the phenomenal concept strategy to explain why they seem so. However, they insist that phenomenal properties are nonetheless real and genuinely qualitative. I have already suggested that this position is problematic. If it is not to collapse into illusionism, then it must employ a notion of phenomenality that is stronger than that of quasi-phenomenality. Phenomenal properties must not merely cause representations of phenomenality but have some genuinely 'feely' aspect to them. And it is unclear what this could be. What phenomenal residue is left, once features such as privacy, intrinsicality, and ineffability have been stripped away (Frankish 2012)?

In practice, reductive explanations of phenomenality tend to take a covertly illusionist form. They typically identify phenomenal character with some functional property of experience such as possession of a certain kind of representational content or availability to higher-order representation. But in so far as these identifications are plausible, it is, I suggest, because subjects whose experiences had this functional property would be disposed to judge that their experiences had a qualitative dimension, rather than because their experiences really would have such a dimension. In the case of higher-order perception theory, for example, it may be true that perceptual awareness of the physical vehicles of experience would create the sense that experiences have an intrinsic quality. But this is an explanation of quasi-phenomenal properties, not phenomenal properties. There is a conflation of phenomenality with the representation of phenomenality, and thus of realism with illusionism.

Of course, these objections assume that we are seeking an *explanation* of consciousness. Physicalists can resist illusionist pressures if they are content to accept the existence of an *explanatory gap* between phenomenal properties and their neural substrates (e.g. Levine 2001). Others, however, may prefer an explicable illusion to an inexplicable reality.

It may be objected that illusionism discards one of the major advantages of conservatism, namely that it gives phenomenal properties a causal role. If phenomenal properties are illusory, then they have no causal role after all. Illusionists can reply that they do not deny that phenomenal concepts track causally effective properties; they merely deny the common-sense view of the nature of these properties—that they are qualitative. Or, perhaps more persuasively, they can say that phenomenal properties *are* causally potent, considered as intentional objects. They move us in the same way that ideas, stories, theories, and memes do, by figuring as the objects of our intentional states. In talking of the power of sensation we are talking of the power of certain representational contents.

2.3. For Illusionism

The case for illusionism can also be made in a positive way, appealing to explanatory considerations. If phenomenal consciousness is conceived as non-physical, then, as Chalmers notes, there is a simple argument for its being illusory (Chalmers 1996, pp. 186–7: Chalmers does not endorse the argument, of course, though he acknowledges its force). If people's claims and beliefs about something (God, say, or UFOs) can be fully explained as arising from causes having no connection with the thing itself, then this is a reason for discounting them and regarding the thing as illusory. But it is widely accepted, even by anti-physicalists, that we do not need to appeal to non-physical properties in order to explain our behaviour and the mental processes that cause it, including our assertions and beliefs about our own conscious experiences. Phenomenal zombies would make the same assertions we do about their conscious experiences and about consciousness in general, and with the same conviction, and they would have beliefs on those matters with the same causal and explanatory roles as ours (though, arguably, with different contents). Given this, our claims and beliefs about consciousness afford no evidence for the truth of phenomenal realism, and it is reasonable to regard them as mistaken.

A second argument for illusionism does not depend on the assumption of anti-physicalism. In general, apparent anomalousness is evidence for illusion. If a property resists explanation in physical terms or is detectable only from a certain perspective, then the simplest explanation is that it is illusory. In this light, considerations usually cited in support of a radical approach to consciousness, such as the existence of an explanatory gap, the conceivability of zombies, and the perspectival nature of phenomenal knowledge, afford equal or greater support for illusionism. Given the force of these considerations, if there is even a remote possibility that we are mistaken about the existence of phenomenal consciousness, then there is a strong abductive inference to the conclusion that we are in fact mistaken about it. And there is reason to think that we could be mistaken about it. For our awareness of phenomenal properties would have to be mediated in some way. If the mind is a representational system, then properties must be mentally represented in order to have cognitive, affective, or motivational significance, and phenomenal properties are no exception, regardless of whether they are physical or non-physical. A creature that lacked introspective representations of its phenomenal properties—we might call it a *representational zombie*—would have no cognitive access to its phenomenal properties and would be unable to form beliefs about them, reflect on them, report them, remember them, respond emotionally to them, or act upon them. Its experiences would not be *like anything*, in the second of the senses distinguished earlier.[10] But we have no introspective way of checking the accuracy of our introspective representations, and so cannot rule out the possibility that they are non-veridical. (Indeed, in so far as we can check, through external inspection of our brain states, they appear to be non-veridical; the properties represented do not show up from other perspectives.) For all we know, then, phenomenality might be illusory; and, given its anomalousness, we can abductively infer that it is.

Illusionism has other explanatory advantages too. One is that it permits us to acknowledge both the wonder of phenomenal consciousness and its potency. This is something realists find hard to do. Stressing the magical, non-physical character of phenomenal properties usually means denying them a causal role, while treating them as physical causes means denying that they are as magical as they seem. But if phenomenal properties are intentional objects, a sort of mental fiction, then we need no longer be embarrassed by them. We can acknowledge how magical and unearthly they are *and* how powerfully they affect us, as intentional objects. In this sense, illusionists may claim to take consciousness more seriously than realists do.

Illusionism also offers an attractive perspective on the function of consciousness. If consciousness has the powerful behavioural influence it seems to have, then we should be able to explain it as an adaptive feature. Again, realists find this hard to do. If consciousness is a matter of pure feel, then it is unclear what function it could perform, and many realists, both radical and conservative, see it as little more than a side effect of perceptual processes. But if consciousness involves an illusion, then new possibilities open. Maybe its function is precisely to give us the impression that we have a magical, non-physical inner life. Humphrey has made a powerful case for such a view (1992, 2006, 2011). He proposes that sensations occur when internalized evaluative responses to stimuli ('sentitions') interact with incoming sensory signals to create complex feedback loops, which, when internally monitored, seem to possess otherworldly, phenomenal properties. This internal 'magic show,' Humphrey argues, powerfully affects the creatures that possess it, giving them a new interest in their existence, inducing them to engage more deeply with their environment (onto which they project phenomenal properties), and creating a sense of self, and, in humans, belief in an ego or immaterial soul. These developments, Humphrey argues, were strongly fitness-enhancing, and the magic show has been sculpted by natural selection to promote them. (This is possible since, on Humphrey's view, the mechanisms of sensation are separate from those of perception and can respond to different evolutionary pressures.) Whether or not this account is right (and it has many attractions), it is an excellent illustration of how evolutionary theorizing about consciousness can flourish, once freed from the metaphysical preoccupations of realism.

3. Defending Illusionism

This section responds to some common objections to illusionism. It argues that they serve primarily to highlight the commitments of the illusionist approach and that illusionists can accommodate weakened versions of the intuitions on which they draw.

3.1. Denying the Data

The most basic objection to illusionism is that it denies the data. To be sure, if all that needed to be explained were the detectable marks of phenomenal consciousness—the related judgments, reports, reactions, dispositions, and so on—then it would be more economical to adopt an illusionist view. But—the objection goes—that is not all that needs to be explained; phenomenal consciousness is itself a datum (Chalmers 1996, p. 188). Phenomenal properties are not theoretical posits introduced to explain other data, but are themselves core data.

There is a sense in which illusionists can agree. It is a datum that phenomenal properties exist as intentional objects; our introspective reports define a notional introspective world which is as we take it to be. But illusionists do, of course, deny that phenomenal properties exist in the real world, as properties of brain states. We are strongly disposed to think that their existence is an introspective datum, but all observation statements, including ones about our own minds, are open to revision in the light of theory. Our introspective reports are data for a science of consciousness, but they require interpretation and evaluation, and the best explanation for them may be one that denies their reliability (Dennett 2003, 2007). And, as we have seen, there are strong theoretical reasons to doubt the reliability of our first-person reports about phenomenal consciousness.

If realists are to maintain that phenomenal consciousness is a datum, then they must say that we have a special kind of epistemic access to it, which excludes any possibility of error. And since no causal process could provide such certainty, they must say that this access is not causally mediated. This is indeed what some realists propose. Chalmers holds that we are *directly acquainted* with phenomenal properties (Chalmers 1996, pp. 192–200). He describes acquaintance as 'a basic sort of epistemic relation between a subject and a property' and says that 'whenever a subject has a phenomenal property, the subject is acquainted with that phenomenal property' (2003, p. 250). Experience is in this sense intrinsically epistemic (1996, p. 196).

This view protects the status of phenomenal consciousness as a datum but does so at a high cost. First, acquaintance can have no psychological significance. In order to talk or think about our phenomenal properties, we need to form mental representations of them, and since representational processes are potentially fallible, the certainty conferred by acquaintance could never be communicated, either to others or even to ourselves, considered as cognitive systems. The price of making consciousness a datum is that the datum is psychologically inert. Second, acquaintance theory assumes that the reactions and associations a sensory episode evokes do not affect its feel, since we are not directly acquainted with them or their effects. Yet there is reason to think that our reactions and associations do shape our sense of what our experiences are like (see Dennett 1988, 1991, chapter 12; 2005, chapter 4). (It might be replied that these factors influence our *judgments* about the feel of our experiences, not the feels themselves, but this would open a systematic gap between what our experiences are like and what we think they are like—which is, at the least, counter-intuitive.)

Acquaintance theory also comes with heavy metaphysical baggage. It is hard to see how physical properties could directly reveal themselves to us, so the theory plausibly assumes an anti-physicalist view of phenomenal consciousness. Moreover, it may require an anti-physicalist view of the experiencing subject too. If subjects are complex physical organisms, how can they become directly acquainted with phenomenal properties? When cognitive scientists talk of information being *available to the subject*, they mean that it is globally broadcast, available for the flexible control of thought and action, and so on. But events need to be represented in order to be available to the subject in this sense. Talk of acquaintance supposes a non-psychological subject, which exists prior to representational processes, as opposed to being partially constituted by them.

This brings us back to talk of it being *like something* to be us. As noted earlier, such talk may mean simply that we have an introspective awareness of our experiences, generated by representational mechanisms. We might call this *introspective subjectivity*. Illusionists agree that we have introspective subjectivity, though they hold that it is radically misleading. But 'like something' talk can be understood in a stronger sense, as indicating that we possess a subjective dimension that is not the product of introspective mechanisms but arises simply from our being the things we are. Call this *intrinsic subjectivity*. When theorists talk of our being directly acquainted with phenomenal properties, it is intrinsic subjectivity they have in mind; the properties, and our awareness of them, are simply correlates of our physical constitution. Plausibly, then, taking phenomenal

consciousness as a datum involves positing intrinsic subjectivity.

Intrinsic subjectivity is, however, deeply mysterious. It is a shadowy companion of physical systems, and we could imagine *any* object possessing it, as panpsychists do. (It might be proposed that only beings with a certain physical structure possess intrinsic subjectivity—perhaps only those that implement an information processing system. But this structure does not *explain* their intrinsic subjectivity, and a creature's reports of what its experiences are like will be the product of introspective mechanisms and will thus manifest introspective subjectivity only.) This does not show that the notion of intrinsic subjectivity is incoherent, but it is, I think, a good reason to explore the idea that it is a fiction created by introspective subjectivity.

3.2. No Appearance–Reality Gap

Another common objection to illusionism is that in the case of qualitative states there is no gap between illusion and reality. Something can look like a Penrose triangle without being a Penrose triangle, but an experience that seems to have a greenish phenomenal character really does have a greenish phenomenal character.[11] As Searle puts it, '*where consciousness is concerned the existence of the appearance is the reality*. If it seems to me exactly as if I am having conscious experiences, then I am having conscious experiences' (Searle 1997, p. 112, italics in original).[12]

This is often presented as a crushing objection to illusionism, but it is far from compelling. It turns on what we mean by *seeming to have* a greenish experience. If we mean having an introspective experience with the same phenomenal feel as a greenish experience, then, trivially, there is no distinction between seeming and reality. But of course that is not what illusionists mean. They mean introspectively representing oneself as having a greenish experience, and one can do this without having a greenish experience. The objector may reply that, in order to create the illusion of a greenish experience, the introspective representation would have to employ a greenish mode of presentation, which would itself have an introspectable greenish feel. However, illusionists will simply deny this, arguing that the content of introspective representations is determined by non-phenomenal, causal or functional factors.[13] The objector may say that there is a big difference between merely representing oneself as having a greenish experience in such a way and actually having a greenish experience, but that is just the point at issue. The illusionist claims that when we think we are having a greenish experience we are in fact merely misrepresenting ourselves as having one. That claim may be false, but the no-gap objection does not add anything to the case against it. Of course, this requires some account of the content of the representations involved, and providing this will be a major challenge for the illusionist. But it is an independent requirement, and the no-gap objection does not make it harder to meet.

Another version of the no-gap objection might go as follows.[14] It is incoherent to doubt that experiences are as they seem, since experience reports are already reports of how things seem. I may come to doubt my initial claim that there is a green patch in front of me and retreat to the more cautious claim that there seems to be a green patch, but I cannot coherently retreat from that claim to the claim that there *seems to seem* to be a green patch. The first claim expresses all the epistemic caution that is necessary or possible. There is something right about this. We have no everyday procedure for correcting sincere and attentive experience reports, and we treat them as authoritative. But it does not follow that this authority is epistemic. Being cautious about the external world does not make one authoritative about the internal one, and seeming to see a green patch isn't the same as infallibly introspecting a greenish phenomenal feel. Rather, as Dennett suggests, the authority might be more like that which a storyteller has over their fictions (Dennett 1991, p. 81).[15]

In a strong form, then, the claim that there is no appearance–reality gap for phenomenal properties is not compelling. A weaker version of the claim is, however, both plausible and compatible with illusionism. From the perspective of a representational theory of mind, the difference between seeming to be aware of a certain phenomenal feel and actually being aware of it is that between having a non-veridical introspective representation of the feel and having a veridical one, and, subjectively, this is no difference at all. In this sense, illusionists can agree that there is no appearance–reality gap for consciousness.

3.3. Who is the Audience?

An illusion presupposes an audience. Who is the audience for the illusion of phenomenal consciousness? Illusionists will join Dennett in

dismissing the idea that there is an inner arena (a 'Cartesian theatre') where perceptual information is assembled and a phenomenal show presented for an appreciation by an internal observer (Dennett 1991). But aren't they committed to reinstating a Cartesian theatre as an arena where the *illusion* of phenomenality is presented?

Illusionists may be committed (as many theorists are) to the existence of an inner *representer* of some kind: a system, or set of processes, which generates introspective representations of sensory states. But this need not amount to an observer, still less a conscious one. If we do not need an inner observer to appreciate perceptual representations, why should we need one to appreciate introspective ones? As Dennett argues, once the brain has made a discrimination, there is no need for another brain system to remake it, and all the work of appreciation and reaction can be (and ultimately must be) distributed among numerous unintelligent subsystems (ibid.). Similarly, once an introspective representation has been generated, the work of reacting to it—of being impressed by the illusion—can, and must, be distributed across such subsystems. There need be no unified audience for the illusion smaller than the organism as a whole (or at least its central nervous system).

That said, illusionists may posit something like an inner display. Recall Humphrey's proposal that internal feedback loops have been shaped by evolution to create a life-enhancing internal magic show (Humphrey 2011). Such a show is, however, different from the one in the Cartesian theatre. First, it is not a phenomenal show, though it is represented as one. Second, it is not a redundant re-presentation of information already encoded in the system. The feedback loops are new features, continuously generated, which need to be monitored and represented in order to have psychological effects. Third, the detector system need do no more than generate representations; again, all the work of appreciating and reacting to the show can be parcelled out to subsystems. Finally, (though Humphrey might not agree) the show need not be a single, integrated one, generating a definitive stream of introspective representations. Instead, there might be numerous micropresentations, yielding multiple drafts of sensation (an introspective counterpart of the multiple drafts model of perceptual processing Dennett proposes; Dennett 1991). Extending the theatrical metaphor, there might be a host of fringe events around the town rather than an official show in a central auditorium.

3.4. Representing Phenomenality

Another objection centres on the representation of phenomenality. If there are no phenomenal properties, how do we represent them? How do we acquire phenomenal concepts, and how do these concepts capture the richness of phenomenality? These are central questions for illusionists, and answering them would go a long way towards solving the illusion problem. Here I shall merely make some preliminary remarks and indicate some lines open to the illusionist.

The task of constructing a theory of content for phenomenal concepts is a difficult one, but it is not obvious that it is significantly more difficult for those who hold that these concepts lack referents. Levine questions whether we can explain the richness and determinacy of our phenomenal representations without reference to actual phenomenal properties (Levine 2001, pp. 146–7). When we think about what an experience is like, he suggests, the phenomenal property itself is included in the thought and serves as its own mode of presentation (ibid., p. 8). The idea that phenomenal concepts quote or incorporate tokens of their referents has been proposed by several theorists (e.g. Chalmers 2003; Papineau 2002, pp. 116–25). However, its explanatory power is questionable. Why should incorporating a phenomenal feel into a representational vehicle make the vehicle represent the feel, let alone in a rich and determinate way? (Incorporating iron filings into it wouldn't make it represent iron.) As Rey stresses, some mechanism would be needed to read off features of the incorporated property and represent them to the rest of the system (Rey 2007, pp. 128–9). But then a secondary, non-quotational account of phenomenal representation would be needed, to which the illusionist could appeal directly.

It is true that illusionism does not sit well with strong externalist views, on which the content of a representation is constituted by causal connections to its referent. Illusionists might argue that phenomenal concepts are compounded from more primitive ones that do refer, or that they have counterfactual causal connections to uninstantiated phenomenal properties. However, there are reasons for finding neither of these options attractive, either for phenomenal concepts or for non-referring concepts generally (Rey 2005). A better option may be to adopt some form of functional-role semantics for phenomenal concepts, on which their content is fixed by their role in mental processing, including their connections to

other concepts, to non-conceptual sensory and introspective representations (their own content determined causally or functionally), and to associations, behavioural dispositions, and so on. (If these functional roles are narrow, 'in the head' ones, the content of our phenomenal representations will be independent of environmental factors—but that is not implausible; see Rey 1998.)

Another possibility is that phenomenal concepts are hybrid ones. Suppose we have a general theoretical concept of a phenomenal property—roughly, that of a simple, intrinsic, immediately known, introspectable property of experience. This concept might be innate, the product of individual theorizing, or culturally acquired. Suppose, too, that we have capacities to introspectively recognize different types of sensory states when they occur, and associated recognitional concepts for the states identified. Then phenomenal concepts might be hybrid ones combining the general theoretical concept with specific recognitional ones. For example, the concept of a certain shade of phenomenal red might be that of *this kind of phenomenal property*, where 'this kind' refers to the kind picked out by the recognitional capacity exercised while having an experience of the relevant type. Of course, if illusionism is true, that capacity does not pick out a phenomenal property; it picks out a complex physical one. So the hybrid concept fails to refer. (Compare 'that kind of ectoplasm' said by a credulous spectator at a séance.) Indeed, the theoretical concept may inform our introspective awareness, so that we mistakenly introspect sensory states *as* phenomenal, just as we might misperceive a flat hologram *as* a three-dimensional object (perhaps even an impossible one, such as a Penrose triangle). A hybrid theory like this may be able to account for many of our intuitions about phenomenal consciousness, rendering illusionism more palatable. If introspection employs recognitional concepts, it may present its objects as being simple, ineffable, and immediately known, but if it is also theoretically informed, it may at the same time radically misrepresent them.

4. Facing up to the Illusion Problem

Illusionism replaces the hard problem with the illusion problem—the problem of explaining how the illusion of phenomenality arises and why it is so powerful. This problem is not easy but not impossibly hard either. The method is to form hypotheses about the underlying cognitive mechanisms and their bases in neurophysiology and neuroanatomy, drawing on evidence from across the cognitive sciences. There are many theoretical options available, and I have indicated some dimensions along which illusionist theories may differ. Some of the issues and positions will be similar to those discussed by conservative realists, but they will assume a new aspect once the commitment to realism is dropped, and we can expect new connections to appear and new theoretical options to present themselves.

Most people find it incredible, even ludicrous, to suppose that phenomenal consciousness is illusory. But if the illusion has been hardwired into our psychology for good evolutionary reasons, then that is to be expected. The question is not whether illusionism is intuitively plausible, but whether it is rationally compelling. If we had a detailed and well-supported illusionist theory, which fully explained our reports, judgments, and intuitions about our own consciousness, would we still want to insist, on reflection, that a hard problem remained? The best way to find out will be to try to construct such a theory.

Our introspective world certainly seems to be painted with rich and potent qualitative properties. But, to adapt James Randi, if Mother Nature is creating that impression by actually equipping our experiences with such properties, then she's doing it the hard way.[16]

REFERENCES

Blackmore, S. *Zen and the Art of Consciousness* (London: Oneworld Publications, 2011).

Carruthers, P. *Phenomenal Consciousness: A Naturalistic Theory* (Cambridge, UK: Cambridge University Press, 2000).

Chalmers, D. J. *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford University Press, 1996).

_____ "The content and epistemology of phenomenal belief," in *Consciousness: New Philosophical Perspectives*, Q. Smith and A. Jokic, eds., (Oxford: Oxford University Press, 2003), pp. 220–72.

Dennett, D. C., "Quining qualia," in *Consciousness in Modern Science*, A. J. Marcel and E. Bisiach, eds., (Oxford: Oxford University Press, 1988), pp. 42–77.

_____*Consciousness Explained* (New York: Little, Brown, 1991).

_____ "Who's on first? Heterophenomenology explained," *Journal of Consciousness Studies* 10, no. 9–10 (2003): pp. 19–30.

_____*Sweet Dreams: Philosophical Obstacles to a Science of Consciousness* (Cambridge, MA: MIT Press, 2005).

_____ "Heterophenomenology reconsidered, *Phenomenology and the Cognitive Sciences* 6, no. 1–2 (2007): pp. 247–70.

Frankish, K., "Quining diet qualia," *Consciousness and Cognition* 21, no. 2 (2012): pp. 667–76.

Hall, R. J., "Phenomenal properties as dummy properties," *Philosophical Studies* 135, no. 2 (2007): pp. 199–223.

Harman, G., :The intrinsic quality of experience," in *Philosophical Perspectives, Vol. 4, Action Theory and Philosophy of Mind*, J. Tomberlin, ed., (Atascadero, CA: Ridgeview. 1990): pp. 31–52.

Hill, C. S., "Imaginability, conceivability, possibility and the mind-body problem," *Philosophical Studies* 87, no. 1 (1997): pp. 61–85.

Hill, C. S., and McLaughlin, B. P., "There are fewer things in reality than are dreamt of in Chalmers's philosophy," *Philosophy and Phenomenological Research* 59, no. 2 (1999): pp. 445–54.

Howell, R., "The Russellian monist's problems with mental causation," *The Philosophical Quarterly* 65, no. 258 (2015): pp. 22–39.

Humphrey, N. *A History of the Mind: Evolution and the Birth of Consciousness* (New York: Simon & Schuster, 1992).

_____*Seeing Red: A Study in Consciousness* (Cambridge, MA: Harvard University Press, 2006).

_____ *Soul Dust: The Magic of Consciousness* (Princeton, NJ: Princeton University Press, 2011).

Kripke, S. A. *Naming and Necessity* (Oxford: Blackwell, 1980).

Levine, J. *Purple Haze: The Puzzle of Consciousness* (Oxford: Oxford University Press, 2001).

Loar, B., "Phenomenal states," in *Philosophical Perspectives, Vol. 4, Action Theory and Philosophy of Mind*, J. E. Tomberlin, ed., (Atascadero, CA: Ridgeview, 1990), pp. 81–108.

Noë, A., ed., *Is the Visual World a Grand Illusion?* (Exeter, UK: Imprint Academic, 2002).

Papineau, D. *Thinking about Consciousness* (Oxford: Oxford University Press, 2002.

Pereboom, D. *Consciousness and the Prospects of Physicalism* (New York: Oxford University Press, 2011).

Place, U. T., "Is consciousness a brain process?" *British Journal of Psychology* 47 (1956): pp. 44–50.

Prinz, J. J., "The fractionation of introspection," *Journal of Consciousness Studies* 11, no. 7–8 (2004): pp. 40–57.

Randi, J., "Science and pseudoscience," in *Carl Sagan's Universe*, Y. Terzian and E. Bilson, eds., (Cambridge, UK: Cambridge University Press, 1997), pp. 170–178.

Rey, G., "Sensational sentences switched," *Philosophical Studies* 68, no. 3 (1992): pp. 289–319.

_____ "Towards a projectivist account of conscious experience," in *Conscious Experience*, T. Metzinger, ed., (Exeter, UK: Imprint Academic, 1995), pp. 123–42.

_____ "A narrow representationalist account of qualitative experience," in *Philosophical Perspectives, Vol. 12, Language, Mind, and Ontology*, J. E. Tomberlin, ed., (Boston, MA: Blackwell, 1998), pp. 435–57

_____ "Philosophical analysis as cognitive psychology: The case of empty concepts," in *Handbook of Categorization in Cognitive Science*, H. Cohen and C. Lefebvre, eds., (Amsterdam: Elsevier, 2005), pp. 71–89

_____"Phenomenal content and the richness and determinacy of colour experience," *Journal of Consciousness Studies* 14, no. 9 (2007): pp. 112–31.

Searle, J. R., *The Mystery of Consciousness*, (New York: New York Review of Books, 1997).

Smart, J. J. C., "Sensations and brain processes," *The Philosophical Review* 68, no. 2 (1959): pp. 141–56.

Strawson, G., "Realistic monism: Why physicalism entails panpsychism," in *Consciousness and its Place in Nature*, A. Freeman, ed., (Exeter, UK: Imprint Academic, 2006), pp. 3–31.

Tartaglia, J., "Conceptualizing physical consciousness," *Philosophical Psychology* 26, no. 6 (2013): pp. 817–38.

Tye, M. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995).

_____ *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000).

Wright, E., "Introduction," in *The Case for Qualia*, E. Wright, ed., (Cambridge, MA: MIT Press, 2008), pp. 1–42.

NOTES

1. When I talk of phenomenal properties not being real or not existing, I mean that they are not *instantiated* in our world. This is compatible with the claim that they exist *qua* properties—a claim which illusionists need not deny.
2. Defenders of illusionist positions (under various names) include Dennett 1988, 1991, 2005; Hall 2007; Humphrey 2011; Pereboom 2011; Rey 1992, 1995, 2007; and Tartaglia 2013. As Tartaglia notes, Place and Smart also denied the existence of phenomenal properties, which Place described as 'mythological' (Place 1956, p. 49; Smart 1959, p. 151).
3. Although all anti-physicalist theories are radical and all conservative theories physicalist, the radical/conservative distinction does not coincide with the

anti-physicalist/ physicalist one, since there may be radical physicalist theories.

4. For an example of a weak illusionist position, see Carruthers 2000, pp. 93–4, 182–91. For more examples, and discussion, see Frankish 2012, where phenomenal properties in this weakly illusionist sense are dubbed *diet qualia*—in contrast to *classic qualia*, or *qualia max*, which are genuinely ineffable, intrinsic, private, and so on. Compare also Levine's distinction between *modest* and *bold* qualophilia (Levine 2001, chapter 5).
5. Extending the soft-drink metaphor, I have dubbed quasi-phenomenal properties *zero qualia* (Frankish 2012).
6. On the varieties of introspection, see Prinz 2004.
7. Pereboom suggests that this might involve the unpacking of a conditional structure in phenomenal concepts (Pereboom 2011, pp. 34–5). As he notes, a given phenomenal property might be reconceptualized either as the neural property that normally causes a representation of the relevant feel or as the higher-order property of being a neural state that could cause a representation of it (ibid.).
8. Compare Dennett's story of the forest god Feenoman (Dennett 1991, chapter 4). Local tribespeople believe Feenoman is real, but visiting anthropologists treat him as an intentional object, defined by the locals' beliefs, and remain neutral on the question of what lies behind the myth. Dennett recommends that we treat first-person phenomenological reports in the same way, as data for third-person theorizing ('heterophenomenology').
9. For the case against first-person science, see Dennett 1991, chapter 4; 2003; 2005, chapter 6; 2007).
10. Compare Rey: 'Postulating qualia properties, whether in the brain or in some special realm, will be of no help unless we have an account of how those properties are assimilated into a person's cognitive life; and it's hard to see how they could be assimilated without being *represented*' (Rey 2007, pp. 129–30).
11. I follow Levine's practice of using 'greenish' for the (putative) feel associated with perception of a green object (Levine 2001).
12. Compare Kripke: 'in the case of mental phenomena there is no 'appearance' beyond the mental phenomenon itself' (Kripke 1980, p. 154).
13. Alternatively, illusionists might concede that introspection employs modes of presentation that appear to have phenomenal feels, but argue that this too is an illusion—that introspection misrepresents the modes of presentation as having phenomenal properties they lack. For defence of this option and an argument that it does not generate an infinite regress, see Pereboom 2011, pp. 27–8.
14. This version was suggested by remarks of Martine Nida-Rümelin, though she might not endorse my presentation of it.
15. It might be argued that phenomenal properties cannot be illusory, since they serve as sense-data, and it is only when sense-data are interpreted that illusion can arise (Wright 2008). This is unpersuasive, however, even granting sense-data theory. Introspective representations of phenomenal properties might serve as data in the construction of representations of external reality while themselves misrepresenting internal, neurophysiological reality. (I am grateful to an anonymous referee for raising this objection.)
16. Earlier versions of this article were presented at the Open University and the University of Crete, and at a 'consciousness cruise' organized by Dmitry Volkoff and the Moscow Center for Consciousness Studies in June 2014, where Jesse Prinz presented a comment on it. My thanks to Jesse and to the audiences on those occasions, mentioning in particular Philip Goff, Martine Nida-Rümelin, Carolyn Price, and Michael Tye. Thanks are also due to Ned Block, Daniel Dennett, Eileen Frankish, Nicholas Humphrey, and Maria Kasmirli for their advice and suggestions. I am especially grateful to David Chalmers for his detailed comments on earlier drafts, from which the article has benefited considerably.

Consciousness and Its Place in Nature

David J. Chalmers

1. Introduction[1]

Consciousness fits uneasily into our conception of the natural world. On the most common conception of nature, the natural world is the physical world. But on the most common conception of consciousness, it is not easy to see how it could be part of the physical world. So it seems that to find a place for consciousness within the natural order, we must either revise our conception of consciousness, or revise our conception of nature.

In twentieth-century philosophy, this dilemma is posed most acutely in C. D. Broad's *The Mind and its Place in Nature* (Broad 1925). The phenomena of mind, for Broad, are the phenomena of consciousness. The central problem is that of locating mind with respect to the physical world. Broad's exhaustive discussion of the problem culminates in a taxonomy of seventeen different views of the mental-physical relation.[2] On Broad's taxonomy, a view might see the mental as nonexistent ('delusive'), as reducible, as emergent, or as a basic property of a substance (a 'differentiating' attribute). The physical might be seen in one of the same four ways. So a four-by-four matrix of views results. (The seventeenth entry arises from Broad's division of the substance/substance view according to whether one substance or two is involved.) At the end, three views are left standing: those on which mentality is an emergent characteristic of either a physical substance or a neutral substance, where in the latter case, the physical might be either emergent or delusive.

In this paper I take my cue from Broad, approaching the problem of consciousness by a strategy of divide-and-conquer. I will not adopt Broad's categories: our understanding of the mind-body problem has advanced in the last 75 years, and it would be nice to think that we have a better understanding of the crucial issues. On my view, the most important views on the metaphysics of consciousness can be divided almost exhaustively into six classes, which I will label 'type A' through 'type F.' Three of these (A through C) involve broadly reductive views, seeing consciousness as a physical process that requires no expansion of a physical ontology. The other three (D through F) involve broadly nonreductive views, on which consciousness involves something irreducible in nature, and requires expansion or reconception of a physical ontology.

The discussion will be cast at an abstract level, giving an overview of the metaphysical landscape. Rather than engaging the empirical science of consciousness, or detailed philosophical theories of consciousness, I will be examining some general classes into which theories of consciousness might fall. I will not pretend to be neutral in this discussion. I think that each of the reductive views is incorrect, while each of the nonreductive views holds some promise. So the first part of this paper can be seen as an extended argument against reductive views of consciousness, while the second part can be seen as an investigation of where we go from there.

2. The Problem

The word 'consciousness' is used in many different ways. It is sometimes used for the ability to discriminate stimuli, or to report information, or to monitor internal states, or to control behavior. We can think of these phenomena as posing the 'easy problems' of consciousness. These are important phenomena, and there is much that is not understood about them, but the problems of explaining them have the character of puzzles rather than mysteries. There seems

From S. Stich and T. Warfield, eds., *Blackwell Guide to the Philosophy of Mind* (Blackwell, 2003).

to be no deep problem in principle with the idea that a physical system could be 'conscious' in these senses, and there is no obvious obstacle to an eventual explanation of these phenomena in neurobiological or computational terms.

The hard problem of consciousness is the problem of experience. Human beings have subjective experience: there is something it is like to be them. We can say that a being is conscious in this sense—or is phenomenally conscious, as it is sometimes put—when there is something it is like to be that being. A mental state is conscious when there is something it is like to be in that state. Conscious states include states of perceptual experience, bodily sensation, mental imagery, emotional experience, occurrent thought, and more. There is something it is like to see a vivid green, to feel a sharp pain, to visualize the Eiffel tower, to feel a deep regret, and to think that one is late. Each of these states has a *phenomenal character,* with *phenomenal properties* (or *qualia*) characterizing what it is like to be in the state.[3]

There is no question that experience is closely associated with physical processes in systems such as brains. It seems that physical processes give rise to experience, at least in the sense that producing a physical system (such as a brain) with the right physical properties inevitably yields corresponding states of experience. But how and why do physical processes give rise to experience? Why do not these processes take place 'in the dark,' without any accompanying states of experience? This is the central mystery of consciousness.

What makes the easy problems easy? For these problems, the task is to explain certain behavioral or cognitive functions: that is, to explain how some causal role is played in the cognitive system, ultimately in the production of behavior. To explain the performance of such a function, one need only specify a mechanism that plays the relevant role. And there is good reason to believe that neural or computational mechanisms can play those roles.

What makes the hard problem hard? Here, the task is not to explain behavioral and cognitive functions: even once one has an explanation of all the relevant functions in the vicinity of consciousness—discrimination, integration, access, report, control—there may still remain a further question: why is the performance of these functions accompanied by experience? Because of this, the hard problem seems to be a different sort of problem, requiring a different sort of solution.

A solution to the hard problem would involve an account of the relation between physical processes and consciousness, explaining on the basis of natural principles how and why it is that physical processes are associated with states of experience. A *reductive explanation* of consciousness will explain this wholly on the basis of physical principles that do not themselves make any appeal to consciousness.[4] A *materialist* (or physicalist) solution will be a solution on which consciousness is itself seen as a physical process. A *nonmaterialist* (or nonphysicalist) solution will be a solution on which consciousness is seen as nonphysical (even if closely associated with physical processes). A *nonreductive* solution will be one on which consciousness (or principles involving consciousness) is admitted as a basic part of the explanation.

It is natural to hope that there will be a materialist solution to the hard problem and a reductive explanation of consciousness, just as there have been reductive explanations of many other phenomena in many other domains. But consciousness seems to resist materialist explanation in a way that other phenomena do not. This resistance can be encapsulated in three related arguments against materialism, summarized in what follows.

3. Arguments against Materialism

3.1. The Explanatory Argument[5]

The first argument is grounded in the difference between the easy problems and the hard problem, as characterized above: the easy problems concern the explanation of behavioral and cognitive functions, but the hard problem does not. One can argue that by the character of physical explanation, physical accounts explain *only* structure and function, where the relevant structures are spatiotemporal structures, and the relevant functions are causal roles in the production of a system's behavior. And one can argue as above that explaining structures and functions does not suffice to explain consciousness. If so, no physical account can explain consciousness.

We can call this the *explanatory argument:*

(1) Physical accounts explain at most structure and function.

(2) Explaining structure and function does not suffice to explain consciousness

---

(3) No physical account can explain consciousness.

If this is right, then while physical accounts can solve the easy problems (which involve only explaining functions), something more is needed to solve the hard problem. It would seem that no reductive explanation of consciousness could succeed. And if we add the premise that what cannot be physically explained is not itself physical (this can be considered an additional final step of the explanatory argument), then materialism about consciousness is false, and the natural world contains more than the physical world.

Of course this sort of argument is controversial. But before examining various ways of responding, it is useful to examine two closely related arguments that also aim to establish that materialism about consciousness is false.

3.2. The Conceivability Argument[6]

According to this argument, it is conceivable that there be a system that is physically identical to a conscious being, but that lacks at least some of that being's conscious states. Such a system might be a *zombie:* a system that is physically identical to a conscious being but that lacks consciousness entirely. It might also be an *invert,* with some of the original being's experiences replaced by different experiences, or a *partial zombie,* with some experiences absent, or a combination thereof. These systems will look identical to a normal conscious being from the third-person perspective: in particular, their brain processes will be molecule-for-molecule identical with the original, and their behavior will be indistinguishable. But things will be different from the first-person point of view. What it is like to be an invert or a partial zombie will differ from what it is like to be the original being. And there is nothing it is like to be a zombie.

There is little reason to believe that zombies exist in the actual world. But many hold that they are at least conceivable: we can coherently imagine zombies, and there is no contradiction in the idea that reveals itself even on reflection. As an extension of the idea, many hold that the same goes for a *zombie world:* a universe physically identical to ours, but in which there is no consciousness. Something similar applies to inverts and other duplicates.

From the conceivability of zombies, proponents of the argument infer their *metaphysical possibility.* Zombies are probably not naturally possible: they probably cannot exist in our world, with its laws of nature. But the argument holds that zombies *could have* existed, perhaps in a very different sort of universe. For example, it is sometimes suggested that God could have created a zombie world, if he had so chosen. From here, it is inferred that consciousness must be nonphysical. If there is a metaphysically possible universe that is physically identical to ours but that lacks consciousness, then consciousness must be a further, nonphysical component of our universe. If God could have created a zombie world, then (as Kripke puts it) after creating the physical processes in our world, he had to do more work to ensure that it contained consciousness.

We can put the argument, in its simplest form, as follows:

(1) It is conceivable that there be zombies
(2) If it is conceivable that there be zombies, it is metaphysically possible that there be zombies.
(3) If it is metaphysically possible that there be zombies, then consciousness is non-physical.

---

(4) Consciousness is nonphysical.

A somewhat more general and precise version of the argument appeals to P, the conjunction of all microphysical truths about the universe, and Q, an arbitrary phenomenal truth about the universe. (Here '∧' represents 'and' and '¬' represents 'not.')

(1) It is conceivable that P∧¬Q.
(2) If it is conceivable that P∧¬Q, it is metaphysically possible that P∧¬Q.
(3) If it is metaphysically possible that P∧¬Q, then materialism is false.

---

(4) Materialism is false.

3.3. The Knowledge Argument[7]

According to the knowledge argument, there are facts about consciousness that are not deducible from physical facts. Someone could know all the physical facts, be a perfect reasoner, and still be unable to know all the facts about consciousness on that basis.

Frank Jackson's canonical version of the argument provides a vivid illustration. On this version, Mary is a neuroscientist who knows everything there is to know about the physical processes relevant to color vision. But Mary has been brought up in a black-and-white room (on an alternative version, she is colorblind[8])

and has never experienced red. Despite all her knowledge, it seems that there is something very important about color vision that Mary does not know: she does not know what it is like to see red. Even complete physical knowledge and unrestricted powers of deduction do not enable her to know this. Later, if she comes to experience red for the first time, she will learn a new fact of which she was previously ignorant: she will learn what it is like to see red.

Jackson's version of the argument can be put as follows (here the premises concern Mary's knowledge when she has not yet experienced red):

(1) Mary knows all the physical facts.
(2) Mary does not know all the facts

---

(3) The physical facts do not exhaust all the facts.

One can put the knowledge argument more generally:

(1) There are truths about consciousness that are not deducible from physical truths.
(2) If there are truths about consciousness that are not deducible from physical truths, then materialism is false.

---

(3) Materialism is false.

3.4. The Shape of the Arguments

These three sorts of argument are closely related. They all start by establishing an *epistemic gap* between the physical and phenomenal domains. Each denies a certain sort of close epistemic relation between the domains: a relation involving what we can know, or conceive, or explain. In particular, each of them denies a certain sort of *epistemic entailment* from physical truths P to the phenomenal truths Q: deducibility of Q from P, or explainability of Q in terms of P, or conceiving of Q upon reflective conceiving of P.

Perhaps the most basic sort of epistemic entailment is a priori entailment, or *implication.* On this notion, P implies Q when the material conditional $P \supset Q$ is a priori; that is, when a subject can know that if P is the case then Q is the case, with justification independent of experience. All of the three arguments above can be seen as making a case against an a priori entailment of Q by P. If a subject who knows only P cannot deduce that Q (as the knowledge argument suggests), or if one can rationally conceive of P without Q (as the conceivability argument suggests), then it seems that P does not imply Q. The explanatory argument can be seen as turning on the claim that an implication from P to Q would require a functional analysis of consciousness, and that the concept of consciousness is not a functional concept.

After establishing an epistemic gap, these arguments proceed by inferring an ontological gap, where ontology concerns the nature of things in the world. The conceivability argument infers from conceivability to metaphysical possibility; the knowledge argument infers from failure of deducibility to difference in facts; and the explanatory argument infers from failure of physical explanation to nonphysicality. One might say that these arguments infer from a failure of epistemic entailment to a failure of ontological entailment. The paradigmatic sort of ontological entailment is *necessitation:* P necessitates Q when the material conditional $P \supset Q$ is metaphysically necessary, or when it is metaphysically impossible for P to hold without Q holding. It is widely agreed that materialism requires that P necessitates all truths (perhaps with minor qualifications). So if there are phenomenal truths Q that P does not necessitate, then materialism is false.

We might call of these arguments *epistemic arguments* against materialism. Epistemic arguments arguably descend from Descartes' arguments against materialism (although these have a slightly different form), and are given their first thorough airing in Broad's book, which contains elements of all three arguments above.[9] The general form of an epistemic argument against materialism is as follows:

(1) There is an epistemic gap between physical and phenomenal truths.
(2) If there is an epistemic gap between physical and phenomenal truths, then there is an ontological gap, and materialism is false.

---

(3) Materialism is false.

Of course this way of looking at things oversimplifies matters, and abstracts away from the differences between the arguments.[10] The same goes for the precise analysis in terms of implication and necessitation. Nevertheless, this analysis provides a useful lens through which to see what the arguments have in common, and through which to analyze various responses to the arguments.

There are roughly three ways that a materialist might resist the epistemic arguments. A type-A materialist denies that there is the relevant sort of epistemic gap. A type-B materialist accepts that there is an unclosable epistemic gap, but denies that there is an ontological gap. And a type-C materialist accepts that there is a deep epistemic gap, but holds that it will eventually be closed. In what follows, I discuss all three of these strategies.

4. Type-A Materialism

According to type-A materialism, there is no epistemic gap between physical and phenomenal truths; or at least, any apparent epistemic gap is easily closed. According to this view, it is not conceivable (at least on reflection) that there be duplicates of conscious beings that have absent or inverted conscious states. On this view, there are no phenomenal truths of which Mary is ignorant in principle from inside her black-and-white room (when she leaves the room, she gains at most an ability). And on this view, on reflection there is no 'hard problem' of explaining consciousness that remains once one has solved the easy problems of explaining the various cognitive, behavioral, and environmental functions.[11]

Type-A materialism sometimes takes the form of eliminativism, holding that consciousness does not exist, and that there are no phenomenal truths. It sometimes takes the form of analytic functionalism or logical behaviorism, holding that consciousness exists, where the concept of 'consciousness' is defined in wholly functional or behavioral terms (e.g., where to be conscious might be to have certain sorts of access to information, and/or certain sorts of dispositions to make verbal reports). For our purposes, the difference between these two views can be seen as terminological. Both agree that we are conscious in the sense of having the functional capacities of access, report, control, and the like; and they agree that we are not conscious in any further (nonfunctionally defined) sense. The analytic functionalist thinks that ordinary terms such as 'conscious' should be used in the first sort of sense (expressing a functional concept), while the eliminativist thinks that it should be used in the second. Beyond this terminological disagreement about the use of existing terms and concepts, the substance of the views is the same.

Some philosophers and scientists who do not explicitly embrace eliminativism, analytic functionalism, and the like are nevertheless recognizably type-A materialists. The characteristic feature of the type-A materialist is the view that on reflection there is nothing in the vicinity of consciousness that needs explaining over and above explaining the various functions: to explain these things is to explain everything in the vicinity that needs to be explained. The relevant functions may be quite subtle and complex, involving fine-grained capacities for access, self-monitoring, report, control, and their interaction, for example. They may also be taken to include all sorts of environmental relations. And the explanation of these functions will probably involve much neurobiological detail. So views that are put forward as rejecting functionalism on the grounds that it neglects biology or neglects the role of the environment may still be type-A views.

One might think that there is room in logical space for a view that denies even this sort of broadly functionalist view of consciousness, but still holds that there is no epistemic gap between physical and phenomenal truths. In practice, there appears to be little room for such a view, for reasons that I will discuss under type C, and there are few examples of such views in practice.[12] So I will take it for granted that a type-A view is one that holds that explaining the functions explains everything, and will class other views that hold that there is no unclosable epistemic gap under type C.

The obvious problem with type-A materialism is that it appears to deny the manifest. It is an uncontested truth that we have the various functional capacities of access, control, report, and the like, and these phenomena pose uncontested explananda (phenomena in need of explanation) for a science of consciousness. But in addition, it seems to be a further truth that we are conscious, and this phenomenon seems to pose a further explanandum. It is this explanandum that raises the interesting problems of consciousness. To flatly deny the further truth, or to deny without argument that there is a hard problem of consciousness over and above the easy problems, would be to make a highly counterintuitive claim that begs the important questions. This is not to say that highly counterintuitive claims are always false, but they need to be supported by extremely strong arguments. So the crucial question is: are there any compelling *arguments* for the claim that on reflection, explaining the functions explains everything?

Type-A materialists often argue by analogy. They point out that in other areas of science,

we accept that explaining the various functions explains the phenomena, so we should accept the same here. In response, an opponent may well accept that in other domains, the functions are all we need to explain. In explaining life, for example, the only phenomena that present themselves as needing explanation are phenomena of adaptation, growth, metabolism, reproduction, and so on, and there is nothing else that even calls out for explanation. But the opponent holds that the case of consciousness is different and possibly unique, precisely because there is something else, phenomenal experience, that calls out for explanation. The type-A materialist must either deny even the appearance of a further explanandum, which seems to deny the obvious, or accept the apparent disanalogy and give further substantial arguments for why, contrary to appearances, only the functions need to be explained.

At this point, type-A materialists often press a different sort of analogy, holding that at various points in the past, thinkers held that there was an analogous epistemic gap for other phenomena, but that these turned out to be physically explained. For example, Dennett 1996 suggests that a vitalist might have held that there was a further 'hard problem' of life over and above explaining the biological function, but that this would have been misguided.

On examining the cases, however, the analogies do not support the type-A materialist. Vitalists typically *accepted,* implicitly or explicitly, that the biological functions in question were what needed explaining. Their vitalism arose because they thought that the functions (adaptation, growth, reproduction, and so on) would not be physically explained. So this is quite different from the case of consciousness. The disanalogy is very clear in the case of Broad. Broad was a vitalist about life, holding that the functions would require a nonmechanical explanation. But at the same time, he held that in the case of life, unlike the case of consciousness, the only evidence we have for the phenomenon is behavioral, and that 'being alive' means exhibiting certain sorts of behavior. Other vitalists were less explicit, but very few of them held that something more than the functions needed explaining (except consciousness itself, in some cases). If a vitalist had held this, the obvious reply would have been that there is no reason to believe in such an explanandum. There is no analogy here.[13]

So these arguments by analogy have no force for the type-A materialist. In other cases, it was always clear that structure and function exhausted the apparent explananda, apart from those tied directly to consciousness itself. So the type-A materialist needs to address the apparent further explanandum in the case of consciousness head on: either flatly denying it, or giving substantial arguments to dissolve it.

Some arguments for type-A materialists proceed indirectly, by pointing out the unsavory metaphysical or epistemological consequences of rejecting the view: e.g., that the rejection leads to dualism, or to problems involving knowledge of consciousness.[14] An opponent will either embrace the consequences or deny that they are consequences. As long as the consequences are not completely untenable, then for the type-A materialist to make progress, this sort of argument needs to be supplemented by a substantial direct argument against the further explanandum.

Such direct arguments are surprisingly hard to find. Many arguments for type-A materialism end up presupposing the conclusion at crucial points. For example, it is sometimes argued (e.g., Rey 1995) that there is no reason to postulate qualia, since they are not needed to explain behavior; but this argument presupposes that only behavior needs explaining. The opponent will hold that qualia are an explanandum in their own right. Similarly, Dennett's use of 'heterophenomenology' (verbal reports) as the primary data to ground his theory of consciousness (Dennett 1991) appears to rest on the assumption that these reports are what need explaining, or that the only 'seemings' that need explaining are dispositions to react and report.

One way to argue for type-A materialism is to argue that there is some intermediate X such that (i) explaining functions suffices to explain X, and (ii) explaining X suffices to explain consciousness. One possible X here is *representation:* it is often held both that conscious states are representational states, representing things in the world, and that we can explain representation in functional terms. If so, it may seem to follow that we can explain consciousness in functional terms. On examination, though, this argument appeals to an ambiguity in the notion of representation. There is a notion of *functional representation,* on which P is represented roughly when a system responds to P and/or produces behavior appropriate for P. In this sense, explaining functioning may explain representation, but explaining representation does not explain consciousness. There is also a notion of *phenomenal representation,*

on which P is represented roughly when a system has a conscious experience as if P. In this sense, explaining representation may explain consciousness, but explaining functioning does not explain representation. Either way, the epistemic gap between the functional and the phenomenal remains as wide as ever. Similar sorts of equivocation can be found with other X's that might be appealed to here, such as 'perception' or 'information.'

Perhaps the most interesting arguments for type-A materialism are those that argue that we can give a physical explanation of our *beliefs* about consciousness, such as the belief that we are conscious, the belief that consciousness is a further explanandum, and the belief that consciousness is nonphysical. From here it is argued that once we have explained the belief, we have done enough to explain, or to explain away, the phenomenon (e.g., Clark 2000, Dennett forthcoming). Here it is worth noting that this only works if the beliefs themselves are functionally analyzable; Chalmers 2002a gives reason to deny this. But even if one accepts that beliefs are ultimately functional, this claim then reduces to the claim that explaining our dispositions to talk about consciousness (and the like) explains everything. An opponent will deny this claim: explaining the dispositions to report may remove the third-person warrant (based on observation of others) for accepting a further explanandum, but it does not remove the crucial first-person warrant (from one's own case). Still, this is a strategy that deserves extended discussion.

At a certain point, the debate between type-A materialists and their opponents usually comes down to intuition: most centrally, the intuition that consciousness (in a nonfunctionally defined sense) exists, or that there is something that needs to be explained (over and above explaining the functions). This claim does not gain its support from argument, but from a sort of observation, along with rebuttal of counterarguments. The intuition appears to be shared by the large majority of philosophers, scientists, and others; and it is so strong that to deny it, a type-A materialist needs exceptionally powerful arguments. The result is that even among materialists, type-A materialists are a distinct minority.

5. Type-B Materialism[15]

According to type-B materialism, there is an epistemic gap between the physical and phenomenal domains, but there is no ontological gap. According to this view, zombies and the like are conceivable, but they are not metaphysically possible. On this view, Mary is ignorant of some phenomenal truths from inside her room, but nevertheless these truths concern an underlying physical reality (when she leaves the room, she learns old facts in a new way). And on this view, while there is a hard problem distinct from the easy problems, it does not correspond to a distinct ontological domain.

The most common form of type-B materialism holds that phenomenal states can be *identified* with certain physical or functional states. This identity is held to be analogous in certain respects (although perhaps not in all respects) with the identity between water and H_2O, or between genes and DNA.[16] These identities are not derived through conceptual analysis, but are discovered empirically: the concept *water* is different from the concept *H_2O*, but they are found to refer to the same thing in nature. On the type-B view, something similar applies to consciousness: the concept of consciousness is distinct from any physical or functional concepts, but we may discover empirically that these refer to the same thing in nature. In this way, we can explain why there is an epistemic gap between the physical and phenomenal domains, while denying any ontological gap. This yields the attractive possibility that we can acknowledge the deep epistemic problems of consciousness while retaining a materialist worldview.

Although such a view is attractive, it faces immediate difficulties. These difficulties stem from the fact that the character of the epistemic gap with consciousness seems to differ from that of epistemic gaps in other domains. For a start, there do not seem to be analogs of the epistemic arguments above in the cases of water, genes, and so on. To explain genes, we merely have to explain why systems function a certain way in transmitting hereditary characteristics; to explain water, we have to explain why a substance has a certain objective structure and behavior. Given a complete physical description of the world, Mary would be able to deduce all the relevant truths about water and about genes, by deducing which systems have the appropriate structure and function. Finally, it seems that we cannot coherently conceive of a world physically identical to our own, in which there is no water, or in which there are no genes. So there is no epistemic gap between the *complete* physical truth about the world and the truth about water and genes that is analogous to the epistemic gap with consciousness.

(Except, perhaps, for epistemic gaps that derive from the epistemic gap for consciousness. For example, perhaps Mary could not deduce or explain the perceptual *appearance* of water from the physical truth about the world. But this would just be another instance of the problem we are concerned with, and so cannot help the type-B materialist.)

So it seems that there is something unique about the case of consciousness. We can put this by saying that while the identity between genes and DNA is empirical, it is not *epistemically primitive:* the identity is itself deducible from the complete physical truth about the world. By contrast, the type-B materialist must hold that the identification between consciousness and physical or functional states is epistemically primitive: the identity is not deducible from the complete physical truth. (If it were deducible, type-A materialism would be true instead.) So the identity between consciousness and a physical state will be a sort of primitive principle in one's theory of the world.

Here, one might suggest that something has gone wrong. Elsewhere, the only sort of place that one finds this sort of primitive principle is in the fundamental laws of physics. Indeed, it is often held that this sort of primitiveness—the inability to be deduced from more basic principles—is the mark of a fundamental law of nature. In effect, the type-B materialist recognizes a principle that has the epistemic status of a fundamental law, but gives it the ontological status of an identity. An opponent will hold that this move is more akin to theft than to honest toil: elsewhere, identifications are grounded in explanations, and primitive principles are acknowledged as fundamental laws.

It is natural to suggest that the same should apply here. If one acknowledges the epistemically primitive connection between physical states and consciousness as a fundamental law, it will follow that consciousness is distinct from any physical property, since fundamental laws always connect distinct properties. So the usual standard will lead to one of the nonreductive views discussed in the second half of this paper. By contrast, the type-B materialist takes an observed connection between physical and phenomenal states, unexplainable in more basic terms, and suggests that it is an identity. This suggestion is made largely in order to preserve a prior commitment to materialism. Unless there is an independent case for primitive identities, the suggestion will seem at best ad hoc and mysterious, and at worst incoherent.

A type-B materialist might respond in various ways. First, some (e.g., Papineau 1993) suggest that identities do not *need* to be explained, so are always primitive. But we have seen that identities in other domains can at least be *deduced* from more basic truths, and so are not primitive in the relevant sense. Second, some (e.g., Block and Stalnaker 1999) suggest that even truths involving water and genes cannot be deduced from underlying physical truths. This matter is too complex to go into here (see Chalmers and Jackson 2001 for a response[17]), but one can note that the epistemic arguments outlined at the beginning suggest a very strong disanalogy between consciousness and other cases. Third, some (e.g., Loar 1990/1997) acknowledge that identities involving consciousness are unlike other identities by being epistemically primitive, but seek to explain this uniqueness by appealing to unique features of the concept of consciousness. This response is perhaps the most interesting, and I will return to it.

There is another line that a type-B materialist can take. One can first note that an *identity* between consciousness and physical states is not strictly required for a materialist position. Rather, one can plausibly hold that materialism about consciousness simply requires that physical states *necessitate* phenomenal states, in that it is metaphysically impossible for the physical states to be present while the phenomenal states are absent or different. That is, materialism requires that entailments $P \supset Q$ be necessary, where P is the complete physical truth about the world and Q is an arbitrary phenomenal truth.

At this point, a type-B materialist can naturally appeal to the work of Kripke 1980, which suggests that some truths are necessarily true without being a priori. For example, Kripke suggests that 'water is H_2O' is necessary—true in all possible worlds—but not knowable a priori. Here, a type-B materialist can suggest that $P \supset Q$ may be a Kripkean a posteriori necessity, like 'water is H_2O' (though it should be noted that Kripke himself denies this claim). If so, then we would *expect* there to be an epistemic gap, since there is no a priori entailment from P to Q, but at the same time there will be no ontological gap. In this way, Kripke's work can seem to be just what the type-B materialist needs.

Here, some of the issues that arose previously arise again. One can argue that in other domains, necessities are not epistemically primitive. The necessary connection between water and H_2O may be a posteriori, but it can itself be deduced from a complete physical description

of the world (one can deduce that water is identical to H_2O, from which it follows that water is necessarily H_2O). The same applies to the other necessities that Kripke discusses. By contrast, the type-B materialist must hold that the connection between physical states and consciousness is epistemically primitive, in that it cannot be deduced from the complete physical truth about the world. Again, one can suggest that this sort of primitive necessary connection is mysterious and ad hoc, and that the connection should instead be viewed as a fundamental law of nature.

I will discuss further problems with these necessities in the next section. But here, it is worth noting that there is a sense in which any type-B materialist position gives up on reductive explanation. Even if type-B materialism is true, we cannot give consciousness the same sort of explanation that we give genes and like, in purely physical terms. Rather, our explanation will always require explanatorily primitive principles to bridge the gap from the physical to the phenomenal. The *explanatory* structure of a theory of consciousness, on such a view, will be very much unlike that of a materialist theory in other domains, and very much like the explanatory structure of the nonreductive theories described below. By labeling these principles identities or necessities rather than laws, the view may preserve the letter of materialism; but by requiring primitive bridging principles, it sacrifices much of materialism's spirit.

6. Type-C Materialism

According to type-C materialism, there is a deep epistemic gap between the physical and phenomenal domains, but it is closable in principle. On this view, zombies and the like are conceivable for us now, but they will not be conceivable in the limit. On this view, it currently seems that Mary lacks information about the phenomenal, but in the limit there would be no information that she lacks. And on this view, while we cannot see now how to solve the hard problem in physical terms, the problem is solvable in principle.

This view is initially very attractive. It seems to acknowledge the deep explanatory gap with which we seem to be faced, while at the same time allowing that the apparent gap may be due to our own limitations. There are different versions of the view. Nagel 1974 has suggested that just as the pre-Socratics could not have understood how matter could be energy, we cannot understand how consciousness could be physical, but a conceptual revolution might allow the relevant understanding. Churchland 1997 suggests that even if we cannot now imagine how consciousness could be a physical process, that is simply a psychological limitation on our part that further progress in science will overcome. Van Gulick 1993 suggests that conceivability arguments are question-begging, since once we have a good explanation of consciousness, zombies and the like will no longer be conceivable. McGinn 1989 has suggested that the problem may be unsolvable by humans due to deep limitations in our cognitive abilities, but that it nevertheless has a solution in principle.

One way to put the view is as follows. Zombies and the like are *prima facie* conceivable (for us now, with our current cognitive processes), but they are not *ideally* conceivable (under idealized rational reflection). Or we could say: phenomenal truths are deducible in principle from physical truths, but the deducibility is akin to that of a complex truth of mathematics: it is accessible in principle (perhaps accessible a priori), but is not accessible to us now, perhaps because the reasoning required is currently beyond us, or perhaps because we do not currently grasp all the required physical truths. If this is so, then there will appear to us that there is a gap between physical processes and consciousness, but there will be no gap in nature.

Despite its appeal, I think that the type-C view is inherently unstable. Upon examination, it turns out either to be untenable, or to collapse into one of the other views on the table. In particular, it seems that the view must collapse into a version of type-A materialism, type-B materialism, type-D dualism, or type-F monism, and so is not ultimately a distinct option.

One way to hold that the epistemic gap might be closed in the limit is to hold that in the limit, we will see that explaining the functions explains everything, and that there is no further explanandum. It is at least coherent to hold that we currently suffer from some sort of conceptual confusion or unclarity that leads us to believe that there is a further explanandum, and that this situation could be cleared up by better reasoning. I will count this position as a version of type-A materialism, not type-C materialism: it is obviously closely related to standard type-A materialism (the main difference is whether we have yet had the relevant insight), and the same issues arise. Like standard type-A

materialism, this view ultimately stands or falls with the strength of (actual and potential) first-order arguments that dissolve any apparent further explanandum.

Once type-A materialism is set aside, the potential options for closing the epistemic gap are highly constrained. These constraints are grounded in the nature of physical concepts, and in the nature of the concept of consciousness. The basic problem has already been mentioned. First: Physical descriptions of the world characterize the world in terms of structure and dynamics. Second: From truths about structure and dynamics, one can deduce only further truths about structure and dynamics. And third: Truths about consciousness are not truths about structure and dynamics. But we can take these steps one at a time.

First: A microphysical description of the world specifies a distribution of particles, fields, and waves in space and time. These basic systems are characterized by their spatiotemporal properties, and properties such as mass, charge, and quantum wavefunction state. These latter properties are ultimately defined in terms of spaces of states that have a certain abstract structure (e.g., the space of continuously varying real quantities, or of Hilbert space states), such that the states play a certain causal role with respect to other states. We can subsume spatiotemporal descriptions and descriptions in terms of properties in these formal spaces under the rubric of *structural* descriptions. The state of these systems can change over time in accord with dynamic principles defined over the relevant properties. The result is a description of the world in terms of its underlying spatiotemporal and formal structure, and dynamic evolution over this structure.

Some type-C materialists hold we do not yet have a complete physics, so we cannot know what such a physics might explain. But here we do not need to have a complete physics: we simply need the claim that physical descriptions are in terms of structure and dynamics. This point is general across physical theories. Such novel theories as relativity, quantum mechanics, and the like may introduce new structures, and new dynamics over those structures, but the general point (and the gap with consciousness) remains.

A type-C materialist might hold that there could be new physical theories that go beyond structure and dynamics. But given the character of physical explanation, it is unclear what sort of theory this could be. Novel physical properties are postulated for their potential in explaining existing physical phenomena, themselves characterized in terms of structure and dynamics, and it seems that structure and dynamics always suffices here. One possibility is that instead of postulating novel properties, physics might end up appealing to consciousness itself, in the way that some theorists hold that quantum mechanics does. This possibility cannot be excluded, but it leads to a view on which consciousness is itself irreducible, and is therefore to be classed in a nonreductive category (type D or type F).

There is one appeal to a 'complete physics' that should be taken seriously. This is the idea that current physics characterizes its underlying properties (such as mass and charge) in terms of abstract structures and relations, but it leaves open their intrinsic natures. On this view, a complete physical description of the world must also characterize the intrinsic properties that ground these structures and relations; and once such intrinsic properties are invoked, physics will go beyond structure and dynamics, in such a way that truths about consciousness may be entailed. The relevant intrinsic properties are unknown to us, but they are knowable in principle. This is an important position, but it is precisely the position discussed under type F, so I defer discussion of it until then.

Second: What can be inferred from this sort of description in terms of structure and dynamics? A low-level microphysical description can entail all sorts of surprising and interesting macroscopic properties, as with the emergence of chemistry from physics, of biology from chemistry, or more generally of complex emergent behaviors in complex systems theory. But in all these cases, the complex properties that are entailed are nevertheless structural and dynamic: they describe complex spatiotemporal structures and complex dynamic patterns of behavior over those structures. So these cases support the general principle that from structure and dynamics, one can infer only structure and dynamics.

A type-C materialist might suggest there are some truths that are not themselves structural-dynamical that are nevertheless implied by a structural-dynamical description. It might be argued, perhaps, that truths about *representation* or *belief* have this character. But as we saw earlier, it seems clear that any sense in which these truths are implied by a structural-dynamic description involves a tacitly functional sense of representation or of belief. This is what we would expect: if claims involving these can

be seen (on conceptual grounds) to be true *in virtue* of a structural-dynamic descriptions holding, the notions involved must themselves be structural-dynamic, at some level.

One might hold that there is some intermediate notion X, such that truths about X hold in virtue of structural-dynamic descriptions, and truths about consciousness hold in virtue of X. But as in the case of type-A materialism, either X is functionally analyzable (in the broad sense), in which case the second step fails, or X is not functionally analyzable, in which case the first step fails. This is brought out clearly in the case of representation: for the notion of functional representation, the first step fails, and for the notion of phenomenal representation, the second step fails. So this sort of strategy can only work by equivocation.

Third: Does explaining or deducing complex structure and dynamics suffice to explain or deduce consciousness? It seems clearly not, for the usual reasons. Mary could know from her black-and-white room all about the spatiotemporal structure and dynamics of the world at all levels, but this will not tell her what it is like to see red. For any complex macroscopic structural or dynamic description of a system, one can conceive of that description being instantiated without consciousness. And explaining structure and dynamics of a human system is only to solve the easy problems, while leaving the hard problems untouched. To resist this last step, an opponent would have to hold that explaining structure and dynamics *thereby* suffices to explain consciousness. The only remotely tenable way to do this would be to embrace type-A materialism, which we have set aside.

A type-C materialist might suggest that instead of leaning on dynamics (as a type-A materialist does), one could lean on structure. Here, spatiotemporal structure seems very unpromising: to explain a system's size, shape, position, motion, and so on is clearly not to explain consciousness. A final possibility is leaning on the structure present in conscious states themselves. Conscious states have structure: there is both internal structure within a single complex conscious state, and there are patterns of similarities and differences between conscious states. But this structure is a distinctively *phenomenal* structure, quite different in kind from the spatiotemporal and formal structure present in physics. The structure of a complex phenomenal state is not spatiotemporal structure (although it may involve the representation of spatiotemporal structure), and the similarities and differences between phenomenal states are not formal similarities and differences, but differences between specific phenomenal characters. This is reflected in the fact that one can conceive of any spatiotemporal structure and formal structure without any associated phenomenal structure; one can know about the first without knowing about the second; and so on. So the epistemic gap is as wide as ever.

The basic problem with any type-C materialist strategy is that epistemic implication from A to B requires some sort of *conceptual hook* by virtue of which the condition described in A can satisfy the conceptual requirements for the truth of B. When a physical account implies truths about life, for example, it does so in virtue of implying information about the macroscopic functioning of physical systems, of the sort required for life: here, broadly functional notions provide the conceptual hook. But in the case of consciousness, no such conceptual hook is available, given the structural-dynamic character of physical concepts, and the quite different character of the concept of consciousness.

Ultimately, it seems that any type-C strategy is doomed for familiar reasons. Once we accept that the concept of consciousness is not itself a functional concept, and that physical descriptions of the world are structural-dynamic descriptions, there is simply no conceptual room for it to be implied by a physical description. So the only room left is to hold that consciousness is a broadly functional concept after all (accepting type-A materialism), hold that there is more in physics than structure and dynamics (accepting type-D dualism or type-F monism), or holding that the truth of materialism does not require an implication from physics to consciousness (accepting type-B materialism).[18] So in the end, there is no separate space for the type-C materialist.

7. Interlude

Are there any other options for the materialist? One further option is to reject the distinctions on which this taxonomy rests. For example, some philosophers, especially followers of Quine 1951, reject any distinction between conceptual truth and empirical truth, or between the a priori and the a posteriori, or between the contingent and the necessary. One who is sufficiently Quinean might therefore reject the distinction between type-A and type-B materialism, holding that talk of epistemic implication

and/or modal entailment is ungrounded, but that materialism is true nevertheless. We might call such a view type-Q materialism. Still, even on this view, similar issues arise. Some Quineans hold that explaining the functions explain everything (Dennett may be an example); if so, all the problems of type-A materialism arise. Others hold that we can postulate identities between physical states and conscious states in virtue of the strong isomorphic connections between them in nature (Paul Churchland may be an example); if so, the problems of type-B materialism arise. Others may appeal to novel future sorts of explanation; if so, the problems of type-C materialism arise. So the Quinean approach cannot avoid the relevant problems.

Leaving this sort of view aside, it looks like the only remotely viable options for the materialist are type-A materialism and type-B materialism. I think that other views are either ultimately unstable, or collapse into one of these (or the three remaining options).[19] It seems to me that the costs of these views—denying the manifest explanandum in the first case, and embracing primitive identities or strong necessities in the second case—suggest very strongly that they are to be avoided unless there are no viable alternatives.

So the residual question is whether there are viable alternatives. If consciousness is not necessitated by physical truths, then it must involve something ontologically novel in the world: to use Kripke's metaphor, after fixing all the physical truths, God had to do more work to fix all the truths about consciousness. That is, there must be ontologically fundamental features of the world over and above the features characterized by physical theory. We are used to the idea that some features of the world are fundamental: in physics, features such as spacetime, mass, and charge, are taken as fundamental and not further explained. If the arguments against materialism are correct, these features from physics do not exhaust the fundamental features of the world: we need to expand our catalog of the world's basic features.

There are two possibilities here. First, it could be that consciousness is itself a fundamental feature of the world, like spacetime and mass. In this case, we can say that phenomenal properties are fundamental. Second, it could be that consciousness is not itself fundamental, but is necessitated by some more primitive fundamental feature X that is not itself necessitated by physics. In this case, we might call X a *protophenomenal* property, and we can say that protophenomenal properties are fundamental. I will typically put things in terms of the first possibility for ease of discussion, but the discussion that follows applies equally to the second. Either way, consciousness involves something novel and fundamental in the world.

The question then arises: how do these novel fundamental properties relate to the already acknowledged fundamental properties of the world, namely those invoked in microphysics? In general, where there are fundamental properties, there are fundamental laws. So we can expect that there will be some sort of fundamental principles—psychophysical laws—connecting physical and phenomenal properties. Like the fundamental laws of relativity or quantum mechanics, these psychophysical laws will not be deducible from more basic principles, but instead will be taken as primitive.

But what is the character of these laws? An immediate worry is that the microphysical aspects of the world are often held to be causally closed, in that every microphysical state has a microphysical sufficient cause. How are fundamental phenomenal properties to be integrated with this causally closed network?

There seem to be three main options for the nonreductionist here. First, one could deny the causal closure of the microphysical, holding that there are causal gaps in microphysical dynamics that are filled by a causal role for distinct phenomenal properties: this is type-D dualism. Second, one could accept the causal closure of the microphysical and hold that phenomenal properties play no causal role with respect to the physical network: this is type-E dualism. Third, one could accept that the microphysical network is causally closed, but hold that phenomenal properties are nevertheless integrated with it and play a causal role, by virtue of constituting the intrinsic nature of the physical: this is type-F monism.

In what follows, I will discuss each of these views. The discussion is necessarily speculative in certain respects, and I do not claim to establish that any one of the views is true or completely unproblematic. But I do aim to suggest that none of them has obvious fatal flaws, and that each deserves further investigation.

8. Type-D Dualism

Type-D dualism holds that microphysics is not causally closed, and that phenomenal properties play a causal role in affecting the physical

world.[20] On this view, usually known as *interactionism*, physical states will cause phenomenal states, and phenomenal states cause physical states. The corresponding psychophysical laws will run in both directions. On this view, the evolution of microphysical states will not be determined by physical principles alone. Psychophysical principles specifying the effect of phenomenal states on physical states will also play an irreducible role.

The most familiar version of this sort of view is Descartes' substance dualism (hence D for Descartes), on which there are separate interacting mental and physical substances or entities. But this sort of view is also compatible with a property dualism, on which there is just one sort of substance or entity with both physical and phenomenal fundamental properties, such that the phenomenal properties play an irreducible role in affecting the physical properties. In particular, the view is compatible with an 'emergentist' view such as Broad's, on which phenomenal properties are ontologically novel properties of physical systems (not deducible from microphysical properties alone), and have novel effects on microphysical properties (not deducible from microphysical principles alone). Such a view would involve basic principles of 'downward' causation of the mental on the microphysical (hence also D for downward causation).

It is sometimes objected that distinct physical and mental states could not interact, since there is no causal nexus between them. But one lesson from Hume and from modern science is that the same goes for any fundamental causal interactions, including those found in physics. Newtonian science reveals no causal nexus by which gravitation works, for example; rather, the relevant laws are simply fundamental. The same goes for basic laws in other physical theories. And the same, presumably, applies to fundamental psychophysical laws: there is no need for a causal nexus distinct from the physical and mental properties themselves.

By far the most influential objection to interactionism is that it is incompatible with physics. It is widely held that science tells us that the microphysical realm is causally closed, so that there is no room for mental states to have any effects. An interactionist might respond in various ways. For example, it could be suggested that although no experimental studies have revealed these effects, none have ruled them out. It might further be suggested that physical theory allows any number of basic *forces* (four as things stand, but there is always room for more), and that an extra force associated with a mental field would be a reasonable extension of existing physical theory. These suggestions would invoke significant revisions to physical theory, so are not to be made lightly; but one could argue that nothing rules them out.

By far the strongest response to this objection is to suggest that far from ruling out interactionism, contemporary physics is positively encouraging to the possibility. On the standard formulation of quantum mechanics, the state of the world is described by a wave function, according to which physical entities are often in a superposed state (e.g., in a superposition of two different positions), even though superpositions are never directly observed. On the standard dynamics, the wave function can evolve in two ways: linear evolution by the Schrödinger equation (which tends to produce superposed states), and nonlinear *collapses* from superposed states into nonsuperposed states. Schrödinger evolution is deterministic, but collapse is nondeterministic. Schrödinger evolution is constantly ongoing, but on the standard formulation, collapses occur only occasionally, on measurement.

The collapse dynamics leaves a door wide open for an interactionist interpretation. Any physical nondeterminism might be held to leave room for nonphysical effects, but the principles of collapse do much more than that. Collapse is supposed to occur on measurement. There is no widely agreed definition of what a measurement is, but there is one sort of event that everyone agrees is a measurement: observation by a conscious observer. Further, it seems that no purely physical criterion for a measurement can work, since purely physical systems are governed by the linear Schrödinger dynamics. As such, it is natural to suggest that a measurement is precisely a conscious observation, and that this conscious observation causes a collapse.

The claim should not be too strong: quantum mechanics does not force this interpretation of the situation onto us, and there are alternative interpretations of quantum mechanics on which there are no collapses, or on which measurement has no special role in collapse.[21] Nevertheless, quantum mechanics appears to be quite *compatible* with such an interpretation. In fact, one might argue that if one was to design elegant laws of physics that allow a role for the conscious mind, one could not do much better than the bipartite dynamics of standard quantum mechanics: one principle governing

deterministic evolution in normal cases, and one principle governing nondeterministic evolution in special situations that have a prima facie link to the mental.

Of course such an interpretation of quantum mechanics is controversial. Many physicists reject it precisely because it is dualistic, giving a fundamental role to consciousness. This rejection is not surprising, but it carries no force when we have independent reason to hold that consciousness may be fundamental. There is some irony in the fact that philosophers reject interactionism on largely physical grounds[22] (it is incompatible with physical theory), while physicists reject an interactionist interpretation of quantum mechanics on largely philosophical grounds (it is dualistic). Taken conjointly, these reasons carry little force, especially in light of the arguments against materialism elsewhere in this paper.

This sort of interpretation needs to be formulated in detail to be assessed.[23] I think the most promising version of such an interpretation allows conscious states to be correlated with the total quantum state of a system, with the extra constraint that conscious states (unlike physical states) can never be superposed. In a conscious physical system such as a brain, the physical and phenomenal states of the system will be correlated in a (nonsuperposed) quantum state. Upon observation of a superposed external system, Schrödinger evolution at the moment of observation would cause the observed system to become correlated with the brain, yielding a resulting superposition of brain states and so (by psychophysical correlation) a superposition of conscious states. But such a superposition cannot occur, so one of the potential resulting conscious states is somehow selected (presumably by a nondeterministic dynamic principle at the phenomenal level). The result is that (by psychophysical correlation) a definite brain state and a definite state of the observed object are also selected. The same might apply to the connection between consciousness and nonconscious processes in the brain: when superposed non-conscious processes threaten to affect consciousness, there will be some sort of selection. In this way, there is a causal role for consciousness in the physical world.

(Interestingly, such a theory may be empirically testable. In quantum mechanics, collapse theories yield predictions slightly different from no-collapse theories, and different hypotheses about the location of collapse yield predictions that differ from each other, although the differences are extremely subtle and are currently impossible to measure. If the relevant experiments can one day be performed, some outcomes would give us strong reason to accept a collapse theory, and might in turn give us grounds to accept a role for consciousness. As a bonus, this could even yield an empirical criterion for the presence of consciousness.)

There are any number of further questions concerning the precise formulation of such a view, its compatibility with physical theory more generally (e.g., relativity and quantum field theory), and its philosophical tenability (e.g., does this view yield the sort of causal role that we are inclined to think consciousness must have?). But at the very least, it cannot be said that physical theory immediately rules out the possibility of an interactionist theory. Those who make this claim often raise their eyebrows when a specific theory such as quantum mechanics is mentioned; but this is quite clearly an inconsistent set of attitudes. If physics is supposed to rule out interactionism, then careful attention to the detail of physical theory is required.

All this suggests that there is at least room for a viable interactionism to be explored, and that the most common objection to interactionism has little force. Of course it does not entail that interactionism is true. There is much that is attractive about the view of the physical world as causally closed, and there is little direct evidence from cognitive science of the hypothesis that behavior cannot be wholly explained in terms of physical causes. Still, if we have independent reason to think that consciousness is irreducible, and if we wish to retain the intuitive view that consciousness plays a causal role, then this is a view to be taken very seriously.

9. Type-E Dualism

Type-E dualism holds that phenomenal properties are ontologically distinct from physical properties, and that the phenomenal has no effect on the physical.[24] This is the view usually known as *epiphenomenalism* (hence type-E): physical states cause phenomenal states, but not vice versa. On this view, psychophysical laws run in one direction only, from physical to phenomenal. The view is naturally combined with the view that the physical realm is causally closed. This further claim is not essential to type-E dualism, but it provides much of the motivation for the view.

As with type-D dualism, type-E dualism is compatible with a substance dualism with distinct physical and mental substances or entities, and is also compatible with a property dualism with one sort of substance or entity and two sorts of properties. Again, it is compatible with an emergentism such as Broad's, on which mental properties are ontologically novel emergent properties of an underlying entity, but in this case although there are emergent qualities, there is no emergent downward causation.

Type-E dualism is usually put forward as respecting both consciousness and science: it simultaneously accommodates the anti-materialist arguments about consciousness and the causal closure of the physical. At the same time, type-E dualism is frequently rejected as deeply counterintuitive. If type-E dualism is correct, then phenomenal states have no effect on our actions, physically construed. For example, a sensation of pain will play no causal role in my hand's moving away from a flame; my experience of decision will play no causal role in my moving to a new country; and a sensation of red will play no causal role in my producing the utterance 'I am experiencing red now.' These consequences are often held to be obviously false, or at least unacceptable.

Still, the type-E dualist can reply that there is no direct *evidence* that contradicts their view. Our evidence reveals only regular connections between phenomenal states and actions, so that certain sorts of experiences are typically followed by certain sorts of actions. Being exposed to this sort of constant conjunction produces a strong *belief* in a causal connection (as Hume pointed out in another context); but it is nevertheless compatible with the absence of a causal connection. Indeed, it seems that if epiphenomenalism *were* true, we would have exactly the same evidence, and be led to believe that consciousness has a causal role for much the same reasons. So if epiphenomenalism is otherwise coherent and acceptable, it seems that these considerations do not provide strong reasons to reject it.[25]

Another objection holds that if consciousness is epiphenomenal, it could not have evolved by natural selection. The type-E dualist has a straightforward reply, however. On the type-E view, there are fundamental psychophysical laws associating physical and phenomenal properties. If evolution selects appropriate physical properties (perhaps involving physical or informational configurations in the brain), then the psychophysical laws will ensure that phenomenal properties are instantiated, too. If the laws have the right form, one can even expect that as more complex physical systems are selected, more complex states of consciousness will evolve. In this way, physical evolution will carry the evolution of consciousness along with it as a sort of byproduct.

Perhaps the most interesting objections to epiphenomenalism focus on the relation between consciousness and representations of consciousness. It is certainly at least strange to suggest that consciousness plays no causal role in my utterances of 'I am conscious.' Some have suggested more strongly that this rules out any *knowledge* of consciousness. It is often held that if a belief about X is to qualify as knowledge, the belief must be caused in some fashion by X. But if consciousness does not effect physical states, and if beliefs are physically constituted, then consciousness cannot cause beliefs. And even if beliefs are not physically constituted, it is not clear how epiphenomenalism can accommodate a causal connection between consciousness and belief.

In response, an epiphenomenalist can deny that knowledge always requires a causal connection. One can argue on independent grounds that there is a stronger connection between consciousness and beliefs about consciousness: consciousness plays a role in *constituting* phenomenal concepts and phenomenal beliefs. A red experience plays a role in constituting a belief that one is having a red experience, for example. If so, there is no causal distance between the experience and the belief. And one can argue that this immediate connection to experience and belief allows for the belief to be justified. If this is right, then epiphenomenalism poses no obstacle to knowledge of consciousness.

A related objection holds that my zombie twin would produce the same reports (e.g., 'I am conscious'), caused by the same mechanisms, and that his reports are unjustified; if so, my own reports are unjustified. In response, one can hold that the true bearers of justification are beliefs, and that my zombie twin and I have *different* beliefs, involving different concepts, because of the role that consciousness plays in constituting my concepts but not the zombie's. Further, the fact that we produce isomorphic reports implies that a third-person observer might not be any more justified in believing that I am conscious than that the zombie is conscious, but it does not imply a difference in first-person justification. The first-person justification for my

belief that I am conscious is not grounded in any way in my reports but rather in my experiences themselves, experiences that the zombie lacks.

I think that there is no knockdown objection to epiphenomenalism here. Still, it must be acknowledged that the situation is at least odd and counterintuitive. The oddness of epiphenomenalism is exacerbated by the fact that the relationship between consciousness and reports about consciousness seems to be something of a lucky coincidence, on the epiphenomenalist view. After all, if psychophysical laws are independent of physical evolution, then there will be possible worlds where physical evolution is the same as ours but the psychophysical laws are very different, so that there is a radical mismatch between reports and experiences. It seems lucky that we are in a world whose psychophysical laws match them up so well. In response, an epiphenomenalist might try to make the case that these laws are somehow the most 'natural' and are to be expected; but there is at least a significant burden of proof here.

Overall, I think that epiphenomenalism is a coherent view without fatal problems. At the same time, it is an inelegant view, producing a fragmented picture of nature, on which physical and phenomenal properties are only very weakly integrated in the natural world. And of course it is a counterintuitive view that many people find difficult to accept. Inelegance and counter-intuitiveness are better than incoherence; so if good arguments force us to epiphenomenalism as the most coherent view, then we should take it seriously. But at the same time, we have good reason to examine other views very carefully.

10. Type-F Monism

Type-F monism is the view that consciousness is constituted by the intrinsic properties of fundamental physical entities: that is, by the categorical bases of fundamental physical dispositions.[26] On this view, phenomenal or protophenomenal properties are located at the fundamental level of physical reality, and in a certain sense, underlie physical reality itself.

This view takes its cue from Bertrand Russell's discussion of physics in *The Analysis of Matter*. Russell pointed out that physics characterizes physical entities and properties by their relations to one another and to us. For example, a quark is characterized by its relations to other physical entities, and a property such as mass is characterized by an associated dispositional role, such as the tendency to resist acceleration. At the same time, physics says nothing about the intrinsic nature of these entities and properties. Where we have relations and dispositions, we expect some underlying intrinsic properties that ground the dispositions, characterizing the entities that stand in these relations.[27] But physics is silent about the intrinsic nature of a quark, or about the intrinsic properties that play the role associated with mass. So this is one metaphysical problem: what are the intrinsic properties of fundamental physical systems?

At the same time, there is another metaphysical problem: how can phenomenal properties be integrated with the physical world? Phenomenal properties seem to be intrinsic properties that are hard to fit in with the structural/dynamic character of physical theory; and arguably, they are the only intrinsic properties that we have direct knowledge of. Russell's insight was that we might solve both these problems at once. Perhaps the intrinsic properties of the physical world are themselves phenomenal properties. Or perhaps the intrinsic properties of the physical world are not phenomenal properties, but nevertheless constitute phenomenal properties: that is, perhaps they are protophenomenal properties. If so, then consciousness and physical reality are deeply intertwined.

This view holds the promise of integrating phenomenal and physical properties very tightly in the natural world. Here, nature consists of entities with intrinsic (proto)phenomenal qualities standing in causal relations within a spacetime manifold. Physics as we know it emerges from the relations between these entities, whereas consciousness as we know it emerges from their intrinsic nature. As a bonus, this view is perfectly compatible with the causal closure of the microphysical, and indeed with existing physical laws. The view can retain the *structure* of physical theory as it already exists; it simply supplements this structure with an intrinsic nature. And the view acknowledges a clear causal role for consciousness in the physical world: (proto)phenomenal properties serve as the ultimate categorical basis of all physical causation.

This view has elements in common with both materialism and dualism. From one perspective, it can be seen as a sort of materialism. If one holds that physical terms refer not to dispositional properties but the underlying

intrinsic properties, then the protophenomenal properties can be seen as physical properties, thus preserving a sort of materialism. From another perspective, it can be seen as a sort of dualism. The view acknowledges phenomenal or protophenomenal properties as ontologically fundamental, and it retains an underlying duality between structural-dispositional properties (those directly characterized in physical theory) and intrinsic protophenomenal properties (those responsible for consciousness). One might suggest that while the view arguably fits the letter of materialism, it shares the spirit of antimaterialism.

In its protophenomenal form, the view can be seen as a sort of neutral monism: there are underlying neutral properties X (the protophenomenal properties), such that the X properties are simultaneously responsible for constituting the physical domain (by their relations) and the phenomenal domain (by their intrinsic nature). In its phenomenal form, the view can be seen as a sort of idealism, such that mental properties constitute physical properties, although these need not be mental properties in the mind of an observer, and they may need to be supplemented by causal and spatiotemporal properties in addition. One could also characterize this form of the view as a sort of panpsychism, with phenomenal properties ubiquitous at the fundamental level. One could give the view in its most general form the name *panprotopsychism*, with either protophenomenal or phenomenal properties underlying all of physical reality.

A type-F monist may have one of a number of attitudes to the zombie argument against materialism. Some type-F monists may hold that a complete physical description must be expanded to include an intrinsic description, and may consequently deny that zombies are conceivable. (We think we are conceiving of a physically identical system only because we overlook intrinsic properties.) Others could maintain that existing physical concepts refer via dispositions to those intrinsic properties that ground the dispositions. If so, these concepts have different primary and secondary intensions, and a type-F monist could correspondingly accept conceivability but deny possibility: we misdescribe the conceived world as physically identical to ours, when in fact it is just structurally identical.[28] Finally, a type-F monist might hold that physical concepts refer to dispositional properties, so that zombies are both conceivable and possible, and the intrinsic properties are not physical properties. The differences between these three attitudes seem to be ultimately terminological rather than substantive.

As for the knowledge argument, a type-F monist might insist that for Mary to have complete physical knowledge, she would have to have a description of the world involving concepts that directly characterize the intrinsic properties; if she had this (as opposed to her impoverished description involving dispositional concepts), she might thereby be in a position to know what it is like to see red. Regarding the explanatory argument, a type-F monist might hold that physical accounts involving intrinsic properties can explain more than structure and function. Alternatively, a type-F monist who sticks to dispositional physical concepts will make responses analogous to one of the other two responses above.

The type-F view is admittedly speculative, and it can sound strange at first hearing. Many find it extremely counterintuitive to suppose that fundamental physical systems have phenomenal properties: e.g., that there is something it is like to be an electron. The protophenomenal version of the view rejects this claim, but retains something of its strangeness: it seems that any properties responsible for constituting consciousness must be strange and unusual properties, of a sort that we might not expect to find in microphysical reality. Still, it is not clear that this strangeness yields any strong objections. Like epiphenomenalism, the view appears to be compatible with all our evidence, and there is no direct evidence against it. One can argue that if the view were true, things would appear to us just as they in fact appear. And we have learned from modern physics that the world is a strange place: we cannot expect it to obey all the dictates of common sense.

One might also object that we do not have any conception of what protophenomenal properties might be like, or of how they could constitute phenomenal properties. This is true, but one could suggest that this is merely a product of our ignorance. In the case of familiar physical properties, there were principled reasons (based on the character of physical concepts) for denying a constitutive connection to phenomenal properties. Here, there are no such principled reasons. At most, there is ignorance of a connection. Of course it would be very desirable to form a positive conception of protophenomenal properties. Perhaps we can do this indirectly, by some sort of theoretical inference from the character of phenomenal properties to their

underlying constituents; or perhaps knowledge of the nature of protophenomenal properties will remain beyond us. Either way, this is no reason to reject the truth of the view.[29]

There is one sort of principled problem in the vicinity, pointed out by James 1890. Our phenomenology has a rich and specific structure: it is unified, bounded, differentiated into many different aspects, but with an underlying homogeneity to many of the aspects, and appears to have a single subject of experience. It is not easy to see how a distribution of a large number of individual microphysical systems, each with their own protophenomenal properties, could somehow add up to this rich and specific structure. Should one not expect something more like a disunified, jagged collection of phenomenal spikes?

This is a version of the *combination problem* for panpsychism (Seager 1995), or what Stoljar 2001 calls the *structural mismatch* problem for the Russellian view (see also Foster 1991, pp. 119–30). To answer it, it seems that we need a much better understanding of the *compositional* principles of phenomenology: that is, the principles by which phenomenal properties can be composed or constituted from underlying phenomenal properties, or protophenomenal properties. We have a good understanding of the principles of physical composition, but no real understanding of the principles of phenomenal composition. This is an area that deserves much close attention: I think it is easily the most serious problem for the type-F monist view. At this point, it is an open question whether or not the problem can be solved.

Some type-F monists appear to hold that they can avoid the combination problem by holding that phenomenal properties are the intrinsic properties of *high-level* physical dispositions (e.g., those involved in neural states), and need not be constituted by the intrinsic properties of microphysical states (hence they may also deny panprotopsychism). But this seems to be untenable: if the low-level network is causally closed and the high-level intrinsic properties are not constituted by low-level intrinsic properties, the high-level intrinsic properties will be epiphenomenal all over again, for familiar reasons. The only way to embrace this position would seem to be in combination with a denial of microphysical causal closure, holding that there are fundamental dispositions above the microphysical level, which have phenomenal properties as their grounds. But such a view would be indistinguishable from type-D dualism.[30] So a distinctive type-F monism will have to face the combination problem directly.

Overall, type-F monism promises a deeply integrated and elegant view of nature. No-one has yet developed a precise and detailed theory in this class, and it is not yet clear whether such a theory can be developed. But at the same time, there appear to be no strong reasons to reject the view. As such, type-F monism is likely to provide fertile grounds for further investigation, and it may ultimately provide the best integration of the physical and the phenomenal within the natural world.

11. Conclusions

Are there any other options for the nonreductionist? There are two views that may not fit straightforwardly into the categories above.

First, some nonmaterialists hold that phenomenal properties are ontologically wholly distinct from physical properties, that microphysics is causally closed, but that phenomenal properties play a causal role with respect to the physical nevertheless. One way this might happen is by a sort of causal overdetermination: physical states causally determine behavior, but phenomenal states cause behavior at the same time. Another is by causal mediation: it might be that in at least some instances of microphysical causation from physical state P_1 to P_2, there is actually a causal connection from P_1 to the mind to P_2, so that the mind enters the causal nexus without altering the structure of the network. And there may be further strategies here. We might call this class type-O dualism (taking overdetermination as a paradigm case). These views share much of the structure of the type-E view (causally closed physical world, distinct phenomenal properties), but escapes the charge of epiphenomenalism. The special causal setups of these views may be hard to swallow, and they share some of the same problems as the type-E view (e.g., the fragmented view of nature, and the 'lucky' psychophysical laws), but this class should nevertheless be put on the table as an option.[31]

Second, some nonmaterialists are *idealists* (in a Berkeleyan sense), holding that the physical world is itself constituted by the conscious states of an observing agent. We might call this view type-I monism. It shares with type-F monism the property that phenomenal states play a role in constituting physical reality, but on the type-I view this happens in a very different

way: not by having separate 'microscopic' phenomenal states underlying each physical state, but rather by having physical states constituted holistically by a 'macroscopic' phenomenal mind. This view seems to be non-naturalistic in a much deeper sense than any of the views above, and in particular seems to suffer from an absence of causal or explanatory closure in nature: once the natural explanation in terms of the external world is removed, highly complex regularities among phenomenal states have to be taken as unexplained in terms of simpler principles. But again, this sort of view should at least be acknowledged.

As I see things, the best options for a nonreductionist are type-D dualism, type-E dualism, or type-F monism: that is, interactionism, epiphenomenalism, or panprotopsychism. If we acknowledge the epistemic gap between the physical and the phenomenal, and we rule out primitive identities and strong necessities, then we are led to a disjunction of these three views. Each of the views has at least some promise, and none have clear fatal flaws. For my part, I give some credence to each of them. I think that in some ways the type-F view is the most appealing, but this sense is largely grounded in aesthetic considerations whose force is unclear.

The choice between these three views may depend in large part on the development of specific theories within these frameworks. Especially for the type-D view and type-F view, further theoretical work is crucial in assessing the theories (e.g., in explicating quantum interactionism, or in understanding phenomenal composition). It may also be that the empirical science of consciousness will give some guidance. As the science progress, we will be led to infer simple principles that underlie correlations between physical and phenomenal states. It may be that these principles turn out to point strongly toward one or the other of these views: e.g., if simple principles connecting microphysical states to phenomenal or protophenomenal states can do the explanatory work, then we may have reason to favor a type-F view, while if the principles latch onto the physical world at a higher level, then we may have reason to favor a type-D or type-E view. And if consciousness has a specific pattern of effects on the physical world, as the type-D view suggests, then empirical studies ought in principle to be able to find these effects, although perhaps only with great difficulty.

Not everyone will agree that each of these views is viable. It may be that further examination will reveal deep problems with some of these views. But this further examination needs to be performed. There has been little critical examination of type-F views to date, for example; we have seen that the standard arguments against type-D views carry very little weight; and while arguments against type-E views carry some intuitive force, they are far from making a knockdown case against the views. I suspect that even if further examination reveals deep problems for some views in this vicinity, it is very unlikely that all such views will be eliminated.

In any case, this gives us some perspective on the mind–body problem. It is often held that even though it is hard to see how materialism could be true, materialism *must* be true, since the alternatives are unacceptable. As I see it, there are at least three prima facie acceptable alternatives to materialism on the table, each of which is compatible with a broadly naturalistic (even if not materialistic) worldview, and none of which has fatal problems. So given the clear arguments against materialism, it seems to me that we should at least tentatively embrace the conclusion that one of these views is correct. Of course all of the views discussed in this paper need to be developed in much more detail, and examined in light of all relevant scientific and philosophical developments, in order to be comprehensively assessed. But as things stand, I think that we have good reason to suppose that consciousness has a fundamental place in nature.

BIBLIOGRAPHY

Armstrong, D. M. *A Materialist Theory of the Mind* (London: Routledge, 1968).

Albert, D. Z. *Quantum Mechanics and Experience* (Cambridge, MA: Harvard University Press, 1993).

Bealer, G., "Mental properties," *Journal of Philosophy* 91 (1994): pp. 185–208.

_____ "Mental causation," (forthcoming)

Block, N., and Stalnaker, R., "Conceptual analysis, dualism, and the explanatory gap," *Philosophical Review* 108 (1999): pp. 1–46.

Broad, C. D. *The Mind and its Place in Nature* (London: Routledge & Kegan Paul, 1925).

Campbell, K. K. *Body and Mind* (New York: Doubleday, 1970).

Carruthers, P. *Phenomenal Consciousness: A Naturalistic Theory* (Cambridge, UK: Cambridge University Press, 2000).

Chalmers, D. J., "Facing up to the problem of consciousness," *Journal of Consciousness Studies* 2 (1995): pp. 200–19. Reprinted in Shear 1997. http://consc.net/papers/facing.html.

_____ *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996).

_____ "Moving forward on the problem of consciousness," *Journal of Consciousness Studies* 4 (1997): pp. 3–46. Reprinted in Shear 1997. http://consc.net/papers/moving.html.

_____ "Materialism and the metaphysics of modality," *Philosophy and Phenomenological Research* 59 (1999): pp. 473–93. http://consc.net/papers/modality.html.

_____ "The content and epistemology of phenomenal belief," in *Consciousness: New Philosophical Essays,* Q. Smith and A. Jokic, eds., (Oxford: Oxford University Press, 2002a). http://consc.net/papers/belief.html.

_____ "Does conceivability entail possibility? "in *Conceivability and Possibility,* T. Gendler and J. Hawthorne, eds., (Oxford University Press, 2002b). http://consc.net/papers/conceivability.html.

Chalmers, D. J., and Jackson, F., "Conceptual analysis and reductive explanation," *Philosophical Review* 110 (2001): pp. 315–61.

Churchland, P. M., "The rediscovery of light," *Journal of Philosophy* 93 (1996): pp. 211–28.

_____ "The hornswoggle problem," 1997. In Shear *Explaining Consciousness: The Hard Problem.* 1997.

Clark, A., "A case where access implies qualia?" *Analysis* 60 (2000): pp. 30–38.

Dennett, D. C. *Consciousness Explained* (Boston: Little-Brown, 1991).

_____ "Facing backward on the problem of consciousness," *Journal of Consciousness Studies* 3 (1996): pp. 4–6.

_____ "The fantasy of first-person science," (forthcoming). http://ase.tufts.edu/cogstud/papers/chalmersdeb3dft.htm.

Dretske, F. *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995).

Evans, G. "Reference and contingency," *The Monist* 62 (1979): pp. 161–89.

Feigl, H., "The 'mental' and the 'physical,'" *Minnesota Studies in the Philosophy of Science* 2 (1958): pp. 370–497. Reprinted (with a postscript) as *The "Mental" and the "Physical"* (Minneapolis: University of Minnesota Press, 1967).

Foster, J. *The Immaterial Self: A Defence of the Cartesian Dualist Conception of the Mind* (New York: Oxford University Press, 1991).

Griffin, D. R. *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem* (Berkeley: University of California Press, 1998).

Harman, G., "The intrinsic quality of experience," *Philosophical Perspectives* 4 (1990): pp. 31–52.

Hill, C. S., "Imaginability, conceivability, possibility, and the mind-body problem," *Philosophical Studies* 87 (1997): pp. 61–85.

Hodgson, D. *The Mind Matters: Consciousness and Choice in a Quantum World* (Oxford University Press, 1991).

Huxley, T., "On the hypothesis that animals are automata, and its history," *Fortnightly Review* 95 (1874): pp. 555–80. Reprinted in *Collected Essays* (London, 1893).

Jackson, F., "A note on physicalism and heat," *Australasian Journal of Philosophy* 58 (1979): pp. 26–34.

_____ "Epiphenomenal qualia," *Philosophical Quarterly* 32 (1982): pp. 127–36.

_____ "Finding the mind in the natural world," in *Philosophy and the Cognitive Sciences,* R. Casati, B. Smith, and G. White, eds., (Vienna: Holder-Pichler-Tempsky, 1994).

James, W. *The Principles of Psychology* (New York: Henry Holt and Co., 1890).

Kaplan, D., "Demonstratives," in *Themes from Kaplan,* J. Almog, J. Perry, and H. Wettstein, eds., (New York: Oxford University Press, 1989).

Kirk, R., "Zombies vs materialists," *Proceedings of the Aristotelian Society (Supplementary Volume)* 48 (1974): pp. 135–52.

Kripke, S. A. *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980).

Levine, J., "Materialism and qualia: The explanatory gap," *Pacific Philosophical Quarterly* 64 (1983): pp. 354–61.

Levine, J. *Purple Haze: The Puzzle of Conscious Experience* (Cambridge, MA: MIT Press, 2000).

Lewis, D., "What experience teaches," in *Proceedings of the Russellian Society* (Sydney, AUS: University of Sydney, 1988).

_____ "Reduction of mind," in *Companion to the Philosophy of Mind,* S. Guttenplan, ed., (Oxford: Blackwell, 1994).

Loar, B., "Phenomenal states," *Philosophical Perspectives* 4 (1990): pp. 81–108. Revised edition in *The Nature of Consciousness,* N. Block, O. Flanagan, and G. Güzeldere, eds., (Cambridge, MA: MIT Press, 1997).

Lockwood, M. *Mind, Brain, and the Quantum* (Oxford University Press, 1989).

Lowe, E. J. *Subjects of Experience* (Cambridge, UK: Cambridge University Press, 1996).

Lycan, W. G. *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996).

Maxwell, G., "Rigid designators and mind-brain identity," *Minnesota Studies in the Philosophy of Science* 9 (1979): pp. 365–403.

Maxwell, N., "Understanding sensations," *Australasian Journal of Philosophy* 46 (1968): pp. 127–45.

McGinn, C., "Can we solve the mind-body problem?" *Mind* 98 (1989): pp. 349–66.

Mills, E., "Interactionism and overdetermination," *American Philosophical Quarterly* 33 (1996): pp. 105–15.

Nagel, T., "What is it like to be a bat?" *Philosophical Review* 83 (1974): pp. 435–50.

Nordby, K., "Vision in a complete achromat: A personal account," in *Night Vision: Basic, Clinical, and Applied Aspects*, R. Hess, L, Sharpe, and K. Nordby, eds., (Cambridge, UK: Cambridge University Press, 1990).

Papineau, D., "Physicalism, consciousness, and the antipathetic fallacy," *Australasian Journal of Philosophy* 71 (1993): pp. 169–83.

Perry, J. *Knowledge, Possibility, and Consciousness* (Cambridge, MA: MIT Press, 2001)>.

Popper, K., and Eccles, J. *The Self and Its Brain: An Argument for Interactionism* (Berlin: Springer, 1977).

Quine, W. V., "Two dogmas of empiricism," *Philosophical Review* 60 (1951): pp. 20–43.

Rey, G., "Toward a projectivist account of conscious experience," in *Conscious Experience*, T. Metzinger, ed., (Paderborn, Germany: Ferdinand Schoningh, 1995).

Robinson, W. S. *Brains and People: An Essay on Mentality and its Causal Conditions* (Philadelphia: Temple University Press, 1998).

Rosenberg, G. H., "*A Place for Consciousness: Probing the Deep Structure of the Natural World*," PhD dissertation, Indiana University, 1997. http://www.ai.uga.edu/~ghrosenb/book.html.

Rosenthal, D. M., "A theory of consciousness," in *The Nature of Consciousness*, N. Block, O. Flanagan, and G. Güzeldere, eds., (Cambridge, MA: MIT Press, 1997).

Russell, B. *The Analysis of Matter* (London: Kegan Paul, 1927).

Ryle, G. *The Concept of Mind* (London: Hutchinson and Co., 1949).

Seager, W., "Consciousness, information and panpsychism," *Journal of Consciousness Studies* 2 (1995): pp. 272–88.

Sellars, W., "Is consciousness physical?" *The Monist* 64 (1981): pp. 66–90.

Shear, J., ed., *Explaining Consciousness: The Hard Problem* (Cambridge, MA: MIT Press, 1997).

Shoemaker, S., "Functionalism and qualia," *Philosophical Studies* 27 (1975): pp. 291–315.

Smart, J. J. C., "Sensations and brain processes," *Philosophical Review* 68 (1959): pp. 141–56.

Stalnaker, R., "Assertion," in *Syntax and Semantics: Pragmatics*, Vol. 9, P. Cole, ed., (New York: Academic Press, 1978).

Stapp, H. *Mind, Matter, and Quantum Mechanics* (Berlin: Springer-Verlag, 1993).

Stoljar, D., "Two conceptions of the physical," *Philosophy and Phenomenological Research* 62 (2001): pp. 253–81.

Strawson, G., "Realistic materialist monism," in *Toward a Science of Consciousness III*, S. Hameroff, A. Kaszniak, and D. Chalmers, eds., (Cambridge, MA: MIT Press, 2000).

Swinburne, R. *The Evolution of the Soul* (New York: Oxford University Press, 1986).

Tye, M. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995).

Van Gulick, R., "Understanding the phenomenal mind: Are we all just armadillos?" in *Consciousness: Philosophical and Psychological Aspects*, M. Davies and G. Humphreys, eds., (Oxford: Blackwell, 1993).

White, S., "Curse of the qualia," *Synthese* 68 (1986): pp. 333–68.

Wigner, E. P., "Remarks on the mind-body question," in *The Scientist Speculates*, I. J. Good, ed., (New York: Basic Books, 1961).

NOTES

1. This paper is an overview of issues concerning the metaphysics of consciousness. Much of the discussion in this paper (especially the first part) recapitulates discussion in Chalmers 1995, 1996, and 1997, although it often takes a different form, and sometimes goes beyond the discussion there. I give a more detailed treatment of many of the issues discussed here in the works cited in the bibliography.
2. The taxonomy is in the final chapter, Chapter 14, of Broad's book (set out on pp. 607–11, and discussed until p. 650). The dramatization of Broad's taxonomy as a 4×4 matrix is illustrated on Andrew Chrucky's website devoted to Broad, at http://www.ditext.com/broad/mpn14.html#t.
3. On my usage, qualia are simply those properties that characterize conscious states according to what it is like to have them. The definition does not build in any further substantive requirements, such as the requirement that qualia are intrinsic or nonintentional. If qualia are intrinsic or nonintentional, this will be a substantive rather than a definitional point (so the claim that the properties of consciousness are nonintrinsic or that they are wholly intentional should not be taken to entail that there are no qualia). Phenomenal properties can also be taken to be properties of individuals (e.g., people) rather than of mental states, characterizing aspects of what it is like to be them at a given time; the difference will not matter much for present purposes.
4. Note that I use 'reductive' in a broader sense than it is sometimes used. Reductive explanation requires only that a high-level phenomenon can be explained wholly in terms of low-level phenomena. This is compatible with the 'multiple realizability' of high-level phenomena in low-level phenomena. For example, there may be many different ways in which digestion could be realized in a physiological system, but one can nevertheless reductively explain a system's digestion in terms of underlying physiology. Another subtlety concerns the possibility of a view on which consciousness can be explained in terms of principles which do not make appeal to consciousness but cannot themselves be physically explained. The definitions above count such a view as neither reductive nor nonreductive. It could reasonably be classified either way, but I

will generally assimilate it with the nonreductive class.

5. A version of the explanatory argument as formulated here is given in Chalmers 1995. For related considerations about explanation, see Levine 1983 on the 'explanatory gap' and Nagel 1974. See also the papers in Shear 1997.
6. Versions of the conceivability argument are put forward by Bealer 1994, Campbell 1970, Chalmers 1996, Kirk 1974, and Kripke 1980, among others. Important predecessors include Descartes' conceivability argument about disembodiment, and Leibniz's 'mill' argument.
7. Sources for the knowledge argument include Jackson 1982, Maxwell 1968, Nagel 1974, and others. Predecessors of the argument are present in Broad's discussion of a 'mathematical archangel' who cannot deduce the smell of ammonia from physical facts (Broad 1925, pp. 70–71), and Feigl's discussion of a 'Martian superscientist' who cannot know what colors look like and what musical tones sound like (Feigl 1958/1967, pp. 64, 68, 140).
8. This version of the thought-experiment has a real-life exemplar in Knut Nordby, a Norwegian sensory biologist who is a rod monochromat (lacking cones in his retina for color vision), and who works on the physiology of color vision. See Nordby 1990.
9. For limited versions of the conceivability argument and the explanatory argument, see Broad, pp. 614–15. For the knowledge argument, see pp. 70–72, where Broad argues that even a 'mathematical archangel' could not deduce the smell of ammonia from microscopic knowledge of atoms. Broad is arguing against 'mechanism,' which is roughly equivalently to contemporary materialism. Perhaps the biggest lacuna in Broad's argument, to contemporary eyes, is any consideration of the possibility that there is an epistemic but not an ontological gap.
10. For a discussion of the relationship between the conceivability argument and the knowledge argument, see Chalmers 1996 and Chalmers 2002b.
11. Type-A materialists include Armstrong 1968, Dennett 1991, Dretske 1995, Harman 1990, Lewis 1988, Rey 1995, and Ryle 1949.
12. Two specific views may be worth mentioning. (i) Some views (e.g., Dretske 1995) deny an epistemic gap while at the same time denying functionalism, by holding that consciousness involves not just functional role but also causal and historical relations to objects in the environment. I count these as type-A views: we can view the relevant relations as part of functional role, broadly construed, and exactly the same considerations arise. (ii) Some views (e.g., Stoljar 2001 and Strawson 2000) deny an epistemic gap not by functionally analyzing consciousness but by expanding our view of the physical base to include underlying intrinsic properties. These views are discussed under type F.
13. In another analogy, Churchland 1996 suggests that someone in Goethe's time might have mounted analogous epistemic arguments against the reductive explanation of 'luminescence.' But on a close look, it is not hard to see that the only further explanandum that could have caused doubts here is the *experience* of seeing light (see Chalmers 1997). This point is no help to the type-A materialist, since this explanandum remains unexplained.
14. For an argument from unsavory metaphysical consequences, see White 1986. For an argument from unsavory epistemological consequences, see Shoemaker 1975. The metaphysical consequences are addressed in the second half of this paper. The epistemological consequences are addressed in Chalmers 2002a.
15. Type-B materialists include Block and Stalnaker 1999, Hill 1997, Levine 1983, Loar 1990/1997, Lycan 1996, Papineau 1993, Perry 2001, and Tye 1995.
16. In certain respects, where type-A materialism can be seen as deriving from the logical behaviorism of Ryle and Carnap, type-B materialism can be seen as deriving from the identity theory of Place and Smart. The matter is complicated, however, by the fact that the early identity-theorists advocated 'topicneutral' (functional) analyses of phenomenal properties, suggesting an underlying type-A materialism.
17. Block and Stalnaker 1999 argue against deducibility in part by arguing that there is usually no explicit conceptual analysis of high-level terms such as 'water' in microphysical terms, or in any other terms that could ground an a priori entailment from microphysical truths to truths about water. In response, Chalmers and Jackson 2001 argue that explicit conceptual analyses are not required for a priori entailments, and that there is good reason to believe that such entailments exist in these cases.
18. Of those mentioned above as apparently sympathetic with type-C materialism, I think McGinn is ultimately a type-F monist, Nagel is either a type-B materialist or a type-F monist, and Churchland is either a type-B materialist or a type-Q materialist (below).
19. One might ask about specific reductive views, such as representationalism (which identifies consciousness with certain representational states), and higher-order thought theory (which identifies consciousness with the objects of higher-order thoughts). How these views are classified depends on how a given theorist regards the representational or higher-order states (e.g., functionally definable or not) and their connection to consciousness (e.g., conceptual or empirical). Among representationalists, I think that Dretske 1995 and Harman 1990 are type-A materialists, while Lycan 1996 and Tye 1995 are type-B materialists. Among higher-order thought theorists, Carruthers 2000 is clearly a type-B materialist, while Rosenthal 1997 is either type-A or type-B. One could also in principle hold nonmaterialist versions of each of these views.
20. Type-D dualists include Foster 1991, Hodgson 1991, Popper and Eccles 1977, Sellars 1981, Stapp 1993, and Swinburne 1986.
21. No-collapse interpretations include Bohm's 'hidden-variable' interpretations, and Everett's 'many-worlds' (or 'many-minds') interpretation. A collapse interpretation that does not invoke measurement is the Ghirardi-Rimini-Weber interpretation (with random occasional collapses). Each of these interpretations requires a significant revision to the standard dynamics of quantum mechanics, and each is controversial, although each has its benefits. (See Albert 1993 for discussion of these and other interpretations.) It is notable that there seems to be no remotely tenable interpretation that preserves the standard claim that collapses occur upon measurement, except for the interpretation involving consciousness.
22. I have been as guilty of this as anyone, setting aside interactionism in Chalmers 1996 partly for reasons

of compatibility with physics. I am still not especially inclined to endorse interactionism, but I now think that the argument from physics is much too glib. Three further reasons for rejecting the view are mentioned in Chalmers 1996. First, if consciousness is to make an interesting qualitative difference to behavior, this requires that it act nonrandomly, in violation of the probabilistic requirements of quantum mechanics. I think there is something to this, but one could bite the bullet on nonrandomness in response, or one could hold that even a random causal role for consciousness is good enough. Second, I argued that denying causal closure yields no special advantage, as a view with causal closure can achieve much the same effect via type-F monism. Again there is something to this, but the type-D view does have the significant advantage of avoiding the type-F view's 'combination problem.' Third, it is not clear that the collapse interpretation yields the *sort* of causal role for consciousness that we expect it to have. I think that this is an important open question that requires detailed investigation.

23. Consciousness-collapse interpretations of quantum mechanics have been put forward by Wigner 1961, Hodgson 1991, and Stapp 1993. Only Stapp goes into much detail, with an interesting but somewhat idiosyncratic account that goes in a direction different from that suggested above.
24. Type-E dualists include Campbell 1970, Huxley 1874, Jackson 1982, and Robinson 1988.
25. Some accuse the epiphenomenalist of a double standard: relying on intuition in making the case against materialism, but going counter to intuition in denying a causal role for consciousness. But intuitions must be assessed against the background of reasons and evidence. To deny the relevant intuitions in the anti-materialist argument (in particular, the intuition of a further explanandum) appears to contradict the available first-person evidence; but denying a causal role for consciousness appears to be compatible on reflection with all our evidence, including first-person evidence.
26. Versions of type-F monism have been put forward by Russell 1927, Feigl 1958/1967, Maxwell 1979, Lockwood 1989, Chalmers 1996, Rosenberg 1997, Griffin 1998, Strawson 2000, and Stoljar 2001.
27. There is philosophical debate over the thesis that all dispositions have a categorical basis. If the thesis is accepted, the case for type-F monism is particularly strong, since microphysical dispositional must have a categorical basis, and we have no independent characterization of that basis. But even if the thesis is rejected, type-F monism is still viable. We need only the thesis that microphysical dispositions *may* have a categorical basis to open room for intrinsic properties here. (Some distinguish intrinsic properties from categorical properties, holding that even dispositional properties are intrinsic. On this view, references to intrinsic properties can be understood as invoking intrinsic categorical properties.)
28. Hence type-F monism is the sort of 'physicalism' that emerges from the loophole mentioned in the two-dimensional argument against type-B materialism. The only way a 'zombie world' W could satisfy the primary intension but not the secondary intension of P is for it to share the dispositional structure of our world but not the underlying intrinsic microphysical properties. If this difference is responsible for the lack of consciousness in W, then the intrinsic microphysical properties in our world are responsible for constituting consciousness. Maxwell 1979 exploits this sort of loophole in replying to Kripke's argument.

    Note that such a W must involve either a different corpus of intrinsic properties from those in our world, or no intrinsic properties at all. A type-F monist who holds that the only coherent intrinsic properties are protophenomenal properties might end up denying the conceivability of zombies, even under a structural-functional description of their physical state—for reasons very different from those of the type-A materialist.
29. McGinn 1991 can be read as advocating a type-F view, while denying that we can know the nature of the protophenomenal properties. His arguments rests on the claim that these properties cannot be known either through perception or through introspection. But this does not rule out the possibility that they might be known through some sort of inference to the best explanation of (introspected) phenomenology, subject to the additional constraints of (perceived) physical structure.
30. In this way, we can see that type-D views and type-F views are quite closely related. We can imagine that if a type-D view is true and there are microphysical causal gaps, we could be led through physical observation alone to postulate higher-level entities to fill these gaps—'psychons,' say—where these are characterized in wholly structural/dispositional terms. The type-D view adds to this the suggestion that psychons have an intrinsic phenomenal nature. The main difference between the type-D view and the type-F view is that the type-D view involves fundamental causation above the microphysical level. This will involve a more radical view of physics, but it might have the advantage of avoiding the combination problem.
31. Type-O positions are advocated by Bealer, (forthcoming), Lowe 1996, and Mills 1996.

Epiphenomenal Qualia

Frank Jackson

It is undeniable that the physical, chemical and biological sciences have provided a great deal of information about the world we live in and about ourselves. I will use the label 'physical information' for this kind of information, and also for information that automatically comes along with it. For example, if a medical scientist tells me enough about the processes that go on in my nervous system, and about how they relate to happenings in the world around me, to what has happened in the past and is likely to happen in the future, to what happens to other similar and dissimilar organisms, and the like, he or she tells me—if I am clever enough to fit it together appropriately—about what is often called the functional role of those states in me (and in organisms in general in similar cases). This information, and its kin, I also label 'physical.'

I do not mean these sketchy remarks to constitute a definition of 'physical information,' and of the correlative notions of physical property, process, and so on, but to indicate what I have in mind here. It is well known that there are problems with giving a precise definition of these notions, and so of the thesis of Physicalism that all (correct) information is physical information.[1] But—unlike some—I take the question of definition to cut across the central problems I want to discuss in this paper.

I am what is sometimes known as a 'qualia freak.' I think that there are certain features of the bodily sensations especially, but also of certain perceptual experiences, which no amount of purely physical information includes. Tell me everything physical there is to tell about what is going on in a living brain, the kind of states, their functional role, their relation to what goes on at other times and in other brains, and so on and so forth, and be I as clever as can be in fitting it all together, you won't have told me about the hurtfulness of pains, the itchiness of itches, pangs of jealousy, or about the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise or seeing the sky.

There are many qualia freaks, and some of them say that their rejection of Physicalism is an unargued intuition.[2] I think that they are being unfair to themselves. They have the following argument. Nothing you could tell of a physical sort captures the smell of a rose, for instance. Therefore, Physicalism is false. By our lights this is a perfectly good argument. It is obviously not to the point to question its validity, and the premise is intuitively obviously true both to them and to me.

I must, however, admit that it is weak from a polemical point of view. There are, unfortunately for us, many who do not find the premise intuitively obvious. The task then is to present an argument whose premises are obvious to all, or at least to as many as possible. This I try to do in §I with what I will call 'the Knowledge argument.' In §II I contrast the Knowledge argument with the Modal argument and in §III with the 'What is it like to be' argument. In §IV I tackle the question of the causal role of qualia. The major factor in stopping people from admitting qualia is the belief that they would have to be given a causal role with respect to the physical world and especially the brain;[3] and it is hard to do this without sounding like someone who believes in fairies. I seek in §IV to turn this objection by arguing that the view that qualia are epiphenomenal is a perfectly possible one.

I. The Knowledge Argument for Qualia

People vary considerably in their ability to discriminate colours. Suppose that in an experiment to catalogue this variation Fred is discovered. Fred has better colour vision than anyone else on record; he makes every discrimination that anyone has ever made, and moreover he makes one that we cannot even begin to make. Show him a batch of ripe tomatoes and he sorts them into two roughly equal groups and does so with complete consistency. That is, if

From *Philosophical Quarterly* 32 (1982): pp. 127–36. Reprinted with permission of the publisher. Addendum excerpted from "What Mary Didn't Know," *Journal of Philosophy* 83 (1986): pp. 291–95, with permission of the author and publisher.

you blindfold him, shuffle the tomatoes up, and then remove the blindfold and ask him to sort them out again, he sorts them into exactly the same two groups.

We ask Fred how he does it. He explains that all ripe tomatoes do not look the same colour to him, and in fact that this is true of a great many objects that we classify together as red. He sees two colours where we see one, and he has in consequence developed for his own use two words 'red_1' and 'red_2' to mark the difference. Perhaps he tells us that he has often tried to teach the difference between red_1 and red_2 to his friends but has got nowhere and has concluded that the rest of the world is red_1-red_2 colour-blind—or perhaps he has had partial success with his children, it doesn't matter. In any case he explains to us that it would be quite wrong to think that because 'red' appears in both 'red_1' and 'red_2' that the two colours are shades of the one colour. He only uses the common term 'red' to fit more easily into our restricted usage. To him red_1 and red_2 are as different from each other and all the other colours as yellow is from blue. And his discriminatory behaviour bears this out: he sorts red_1 from red_2 tomatoes with the greatest of ease in a wide variety of viewing circumstances. Moreover, an investigation of the physiological basis of Fred's exceptional ability reveals that Fred's optical system is able to separate out two groups of wavelengths in the red spectrum as sharply as we are able to sort out yellow from blue.[4]

I think that we should admit that Fred can see, really see, at least one more colour than we can; red_1 is a different colour from red_2. We are to Fred as a totally red-green colour-blind person is to us. H. G. Wells' story 'The Country of the Blind' is about a sighted person in a totally blind community.[5] This person never manages to convince them that he can see, that he has an extra sense. They ridicule this sense as quite inconceivable, and treat his capacity to avoid falling into ditches, to win fights and so on as precisely that capacity and nothing more. We would be making their mistake if we refused to allow that Fred can see one more colour than we can.

What kind of experience does Fred have when he sees red_1 and red_2? What is the new colour or colours like? We would dearly like to know but do not; and it seems that no amount of physical information about Fred's brain and optical system tells us. We find out perhaps that Fred's cones respond differentially to certain light waves in the red section of the spectrum that make no difference to ours (or perhaps he has an extra cone) and that this leads in Fred to a wider range of those brain states responsible for visual discriminatory behaviour. But none of this tells us what we really want to know about his colour experience. There is something about it we don't know. But we know, we may suppose, everything about Fred's body, his behaviour and dispositions to behaviour and about his internal physiology, and everything about his history and relation to others that can be given in physical accounts of persons. We have all the physical information. Therefore, knowing all this is *not* knowing everything about Fred. It follows that Physicalism leaves something out.

To reinforce this conclusion, imagine that as a result of our investigations into the internal workings of Fred we find out how to make everyone's physiology like Fred's in the relevant respects; or perhaps Fred donates his body to science and on his death we are able to transplant his optical system into someone else—again the fine detail doesn't matter. The important point is that such a happening would create enormous interest. People would say, 'At last we will know what it is like to see the extra colour, at last we will know how Fred has differed from us in the way he has struggled to tell us about for so long.' Then it cannot be that we knew all along all about Fred. But *ex hypothesi* we did know all along everything about Fred that features in the physicalist scheme; hence the physicalist scheme leaves something out.

Put it this way. *After* the operation, we will know *more* about Fred and especially about his colour experiences. But beforehand we had all the physical information we could desire about his body and brain, and indeed everything that has ever featured in physicalist accounts of mind and consciousness. Hence there is more to know than all that. Hence Physicalism is incomplete.

Fred and the new colour(s) are of course essentially rhetorical devices. The same point can be made with normal people and familiar colours. Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room *via* a black and white television monitor. She specialises in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red,' 'blue,' and so on. She discovers, for example, just which wave-length combinations from the sky stimulate the retina, and exactly how this produces *via* the central nervous system the

contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue.' (It can hardly be denied that it is in principle possible to obtain all this physical information from black and white television, otherwise the Open University would *of necessity* need to use colour television.)

What will happen when Mary is released from her black and white room or is given a colour television monitor? Will she *learn* anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. Ergo there is more to have than that, and Physicalism is false.

Clearly the same style of Knowledge argument could be deployed for taste, hearing, the bodily sensations and generally speaking for the various mental states which are said to have (as it is variously put) raw feels, phenomenal features or qualia. The conclusion in each case is that the qualia are left out of the physicalist story. And the polemical strength of the Knowledge argument is that it is so hard to deny the central claim that one can have all the physical information without having all the information there is to have.

II. The Modal Argument

By the Modal Argument I mean an argument of the following style.[6] Sceptics about other minds are not making a mistake in deductive logic, whatever else may be wrong with their position. No amount of physical information about another *logically entails* that he or she is conscious or feels anything at all. Consequently there is a possible world with organisms exactly like us in every physical respect (and remember that includes functional states, physical history, et al.) but which differ from us profoundly in that they have no conscious mental life at all. But then what is it that we have and they lack? Not anything physical *ex hypothesi*. In all physical regards we and they are exactly alike. Consequently there is more to us than the purely physical. Thus Physicalism is false.[7]

It is sometimes objected that the Modal argument misconceives Physicalism on the ground that that doctrine is advanced as a *contingent* truth.[8] But to say this is only to say that physicalists restrict their claim to *some* possible worlds, including especially ours; and the Modal argument is only directed against this lesser claim. If we in *our* world, let alone beings in any others, have features additional to those of our physical replicas in other possible worlds, then we have non-physical features or qualia.

The trouble rather with the Modal argument is that it rests on a disputable modal intuition. Disputable because it is disputed. Some sincerely deny that there can be physical replicas of us in other possible worlds which nevertheless lack consciousness. Moreover, at least one person who once had the intuition now has doubts.[9]

Head-counting may seem a poor approach to a discussion of the Modal argument. But frequently we can do no better when modal intuitions are in question, and remember our initial goal was to find the argument with the greatest polemical utility.

Of course, qua protagonists of the Knowledge argument we may well accept the modal intuition in question; but this will be a *consequence* of our already having an argument to the conclusion that qualia are left out of the physicalist story, not our ground for that conclusion. Moreover, the matter is complicated by the possibility that the connection between matters physical and qualia is like that sometimes held to obtain between aesthetic qualities and natural ones. Two possible worlds which agree in all 'natural' respects (including the experiences of sentient creatures) must agree in all aesthetic qualities also, but it is plausibly held that the aesthetic qualities cannot be reduced to the natural.

III. The 'What Is It Like to Be' Argument

In 'What is it like to be a bat?' Thomas Nagel argues that no amount of physical information can tell us what it is like to be a bat, and indeed that we, human beings, cannot imagine what it is like to be a bat.[10] His reason is that what this is like can only be understood from a bat's point of view, which is not our point of view and is not something capturable in physical terms which are essentially terms understandable equally from many points of view.

It is important to distinguish this argument from the Knowledge argument. When I complained that all the physical knowledge about Fred was not enough to tell us what his special colour experience was like, I was not complaining that we weren't finding out what it is like to *be* Fred. I was complaining that there is

something *about* his experience, a property of it, of which we were left ignorant. And if and when we come to know what this property is we still will not know what it is like to *be* Fred, but we will know more *about* him. No amount of knowledge about Fred, be it physical or not, amounts to knowledge 'from the inside' concerning Fred. We are not Fred. There is thus a whole set of items of knowledge expressed by forms of words like 'that it is *I myself* who is . . .' which Fred has and we simply cannot have because we are not him.[11]

When Fred sees the colour he alone can see, one thing he knows is the way his experience of it differs from his experience of seeing red and so on, *another* is that he himself is seeing it. Physicalist and qualia freaks alike should acknowledge that no amount of information of whatever kind that *others* have *about* Fred amounts to knowledge of the second. My complaint though concerned the first and was that the special quality of his experience is certainly a fact about it, and one which Physicalism leaves out because no amount of physical information told us what it is.

Nagel speaks as if the problem he is raising is one of extrapolating from knowledge of one experience to another, of imagining what an unfamiliar experience would be like on the basis of familiar ones. In terms of Hume's example, from knowledge of some shades of blue we can work out what it would be like to see other shades of blue. Nagel argues that the trouble with bats et al. is that they are too unlike us. It is hard to see an objection to Physicalism here. Physicalism makes no special claims about the imaginative or extrapolative powers of human beings, and it is hard to see why it need do so.[12]

Anyway, our Knowledge argument makes no assumptions on this point. If Physicalism were true, enough physical information about Fred would obviate any need to extrapolate or to perform special feats of imagination or understanding in order to know all about his special colour experience. *The information would already be in our possession.* But it clearly isn't. That was the nub of the argument.

IV. The Bogey of Epiphenomenalism

Is there any really *good* reason for refusing to countenance the idea that qualia are causally impotent with respect to the physical world? I will argue for the answer no, but in doing this I will say nothing about two views associated with the classical epiphenomenalist position. The first is that mental *states* are inefficacious with respect to the physical world. All I will be concerned to defend is that it is possible to hold that certain *properties* of certain mental states, namely those I've called qualia, are such that their possession or absence makes no difference to the physical world. The second is that the mental is *totally* causally inefficacious. For all I will say it may be that you have to hold that the instantiation of *qualia* makes a difference to *other mental states* though not to anything physical. Indeed general considerations to do with how you could come to be aware of the instantiation of qualia suggest such a position.[13]

Three reasons are standardly given for holding that a quale like the hurtfulness of a pain must be causally efficacious in the physical world, and so, for instance, that its instantiation must sometimes make a difference to what happens in the brain. None, I will argue, has any real force. (I am much indebted to Alec Hyslop and John Lucas for convincing me of this.)

(i) It is supposed to be just obvious that the hurtfulness of pain is partly responsible for the subject seeking to avoid pain, saying 'It hurts' and so on. But, to reverse Hume, anything can fail to cause anything. No matter how often *B* follows *A*, and no matter how initially obvious the causality of the connection seems, the hypothesis that *A* causes *B* can be overturned by an over-arching theory which shows the two as distinct effects of a common underlying causal process.

To the untutored the image on the screen of Lee Marvin's fist moving from left to right immediately followed by the image of John Wayne's head moving in the same general direction looks as causal as anything.[14] And of course throughout countless Westerns images similar to the first are followed by images similar to the second. All this counts for precisely nothing when we know the over-arching theory concerning how the relevant images are both effects of an underlying causal process involving the projector and the film. The epiphenomenalist can say exactly the same about the connection between, for example, hurtfulness and behaviour. It is simply a consequence of the fact that certain happenings in the brain cause both.

(ii) The second objection relates to Darwin's Theory of Evolution. According to

natural selection the traits that evolve over time are those conducive to physical survival. We may assume that qualia evolved over time—we have them, the earliest forms of life do not—and so we should expect qualia to be conducive to survival. The objection is that they could hardly help us to survive if they do nothing to the physical world.

The appeal of this argument is undeniable, but there is a good reply to it. Polar bears have particularly thick, warm coats. The Theory of Evolution explains this (we suppose) by pointing out that having a thick, warm coat is conducive to survival in the Arctic. But having a thick coat goes along with having a heavy coat, and having a heavy coat is *not* conducive to survival. It slows the animal down.

Does this mean that we have refuted Darwin because we have found an evolved trait—having a heavy coat—which is not conducive to survival? Clearly not. Having a heavy coat is an unavoidable concomitant of having a warm coat (in the context, modern insulation was not available), and the advantages for survival of having a warm coat outweighed the disadvantages of having a heavy one. The point is that all we can extract from Darwin's theory is that we should expect any evolved characteristic to be *either* conducive to survival *or* a by-product of one that is so conducive. The epiphenomenalist holds that qualia fall into the latter category. They are a by-product of certain brain processes that are highly conducive to survival.

(iii) The third objection is based on a point about how we come to know about other minds. We know about other minds by knowing about other behaviour, at least in part. The nature of the inference is a matter of some controversy, but it is not a matter of controversy that it proceeds from behaviour. That is why we think that stones do not feel and dogs do feel. But, runs the objection, how can a person's behaviour provide any reason for believing he has qualia like mine, or indeed any qualia at all, unless this behaviour can be regarded as the *outcome* of the qualia. Man Friday's footprint was evidence of Man Friday because footprints are causal outcomes of feet attached to people. And an epiphenomenalist cannot regard behaviour, or indeed anything physical, as an outcome of qualia.

But consider my reading in *The Times* that Spurs won. This provides excellent evidence that *The Telegraph* has also reported that Spurs won, despite the fact that (I trust) *The Telegraph* does not get the results from *The Times*. They each send their own reporters to the game. *The Telegraph*'s report is in no sense an outcome of *The Times*,' but the latter provides good evidence for the former nevertheless.

The reasoning involved can be reconstructed thus. I read in *The Times* that Spurs won. This gives me reason to think that Spurs won because I know that Spurs' winning is the most likely candidate to be what caused the report in *The Times*. But I also know that Spurs' winning would have had many effects, including almost certainly a report in *The Telegraph*.

I am arguing from one effect back to its cause and out again to another effect. The fact that neither effect causes the other is irrelevant. Now the epiphenomenalist allows that qualia are effects of what goes on in the brain. Qualia cause nothing physical but are caused by something physical. Hence the epiphenomenalist can argue from the behaviour of others to the qualia of others by arguing from the behaviour of others back to its causes in the brains of others and out again to their qualia.

You may well feel for one reason or another that this is a more dubious chain of reasoning than its model in the case of newspaper reports. You are right. The problem of other minds is a major philosophical problem, the problem of other newspaper reports is not. But there is no special problem of Epiphenomenalism as opposed to, say, Interactionism here.

There is a very understandable response to the three replies I have just made. 'All right, there is no knockdown refutation of the existence of epiphenomenal qualia. But the fact remains that they are an excrescence. They *do* nothing, they *explain* nothing, they serve merely to soothe the intuitions of dualists, and it is left a total mystery how they fit into the world view of science. In short we do not and cannot understand the how and why of them.'

This is perfectly true; but is no objection to qualia, for it rests on an overly optimistic view of the human animal, and its powers. We are the products of Evolution. We understand and sense what we need to understand and sense in order to survive. Epiphenomenal qualia are totally irrelevant to survival. At no stage of our evolution did natural selection favour those who could make sense of how they are caused

and the laws governing them, or in fact why they exist at all. And that is why we can't.

It is not sufficiently appreciated that Physicalism is an extremely optimistic view of our powers. If it is true, we have, in very broad outline admittedly, a grasp of our place in the scheme of things. Certain matters of sheer complexity defeat us—there are an awful lot of neurons—but in principle we have it all. But consider the antecedent probability that everything in the Universe be of a kind that is relevant in some way or other to the survival of *homo sapiens*. It is very low surely. But then one must admit that it is very likely that there is a part of the whole scheme of things, maybe a big part, which no amount of evolution will ever bring us near to knowledge about or understanding. For the simple reason that such knowledge and understanding is irrelevant to survival.

Physicalists typically emphasise that we are a part of nature on their view, which is fair enough. But if we are a part of nature, we are as nature has left us after however many years of evolution it is, and each step in that evolutionary progression has been a matter of chance constrained just by the need to preserve or increase survival value. The wonder is that we understand as much as we do, and there is no wonder that there should be matters which fall quite outside our comprehension. Perhaps exactly how epiphenomenal qualia fit into the scheme of things is one such.

This may seem an unduly pessimistic view of our capacity to articulate a truly comprehensive picture of our world and our place in it. But suppose we discovered living on the bottom of the deepest oceans a sort of sea slug which manifested intelligence. Perhaps survival in the conditions required rational powers. Despite their intelligence, these sea slugs have only a very restricted conception of the world by comparison with ours, the explanation for this being the nature of their immediate environment. Nevertheless they have developed sciences which work surprisingly well in these restricted terms. They also have philosophers, called slugists. Some call themselves tough-minded slugists, others confess to being soft-minded slugists.

The tough-minded slugists hold that the restricted terms (or ones pretty like them which may be introduced as their sciences progress) suffice in principle to describe everything without remainder. These tough-minded slugists admit in moments of weakness to a feeling that their theory leaves something out. They resist this feeling and their opponents, the soft-minded slugists, by pointing out—absolutely correctly—that no slugist has ever succeeded in spelling out how this mysterious residue fits into the highly successful view that their sciences have and are developing of how their world works.

Our sea slugs don't exist, but they might. And there might also exist super beings which stand to us as we stand to the sea slugs. We cannot adopt the perspective of these super beings, because we are not them, but the possibility of such a perspective is, I think, an antidote to excessive optimism.[15]

ADDENDUM: FROM 'WHAT MARY DIDN'T KNOW'

I. Three Clarifications

The knowledge argument does not rest on the dubious claim that logically you cannot imagine what sensing red is like unless you have sensed red. Powers of imagination are not to the point. The contention about Mary is not that, despite her fantastic grasp of neurophysiology and everything else physical, she *could not imagine* what it is like to sense red; it is that, as a matter of fact, she *would not know*. But if physicalism is true, she would know; and no great powers of imagination would be called for. Imagination is a faculty that those who *lack* knowledge need to fall back on.

Secondly, the intensionality of knowledge is not to the point. The argument does not rest on assuming falsely that, if S knows that a is F and if $a = b$, then S knows that b is F. It is concerned with the nature of Mary's total body of knowledge before she is released: is it complete, or do some facts escape it? What is to the point is that S may know that a is F and *know* that $a = b$, yet arguably not know that b is F, by virtue of not being sufficiently logically alert to follow the consequences through. If Mary's lack of knowledge were at all like this, there would be no threat to physicalism in it. But it is very hard to believe that her lack of knowledge could be remedied merely by her explicitly following through enough logical consequences of her vast physical knowledge. Endowing her with great logical acumen and persistence is not in itself enough to fill in the gaps in her

knowledge. On being let out, she will not say 'I could have worked all this out before by making some more purely logical inferences.'

Thirdly, the knowledge Mary lacked which is of particular point for the knowledge argument against physicalism is *knowledge about the experiences of others,* not about her own. When she is let out, she has new experiences, color experiences she has never had before. It is not, therefore, an objection to physicalism that she learns *something* on being let out. Before she was let out, she could not have known facts about her experience of red, for there were no such facts to know. That physicalist and non-physicalist alike can agree on. After she is let out, things change; and physicalism can happily admit that she learns this; after all, some physical things will change, for instance, her brain states and their functional roles. The trouble for physicalism is that, after Mary sees her first ripe tomato, she will realize how impoverished her conception of the mental life of *others* has been *all along.* She will realize that there was, all the time she was carrying out her laborious investigations into the neurophysiologies of others and into the functional roles of their internal states, something about these people she was quite unaware of. All along their experiences (or many of them, those got from tomatoes, the sky, . . .) had a feature conspicuous to them but until now hidden from her (in fact, not in logic). But she knew all the physical facts about them all along; hence, what she did not know until her release is not a physical fact about their experiences. But it is a fact about them. That is the trouble for physicalism.

II. Churchland's Three Objections[16]

(i) Churchland's first objection is that the knowledge argument contains a defect that 'is simplicity itself' (23). The argument equivocates on the sense of 'knows about.' How so? Churchland suggests that the following is 'a conveniently tightened version' of the knowledge argument:

(1) Mary knows everything there is to know about brain states and their properties.
(2) It is not the case that Mary knows everything there is to know about sensations and their properties.

Therefore, by Leibniz's law,

(3) Sensations and their properties ≠ brain states and their properties (23).

Churchland observes, plausibly enough, that the type or kind of knowledge involved in premise 1 is distinct from the kind of knowledge involved in premise 2. We might follow his lead and tag the first 'knowledge by description,' and the second 'knowledge by acquaintance'; but, whatever the tags, he is right that the displayed argument involves a highly dubious use of Leibniz's law.

My reply is that the displayed argument may be convenient, but it is not accurate. It is not the knowledge argument. Take, for instance, premise 1. The whole thrust of the knowledge argument is that Mary (before her release) does *not* know everything there is to know about brain states and their properties, because she does not know about certain qualia associated with them. What is complete, according to the argument, is her knowledge of matters physical. A convenient and accurate way of displaying the argument is:

(1)′ Mary (before her release) knows everything physical there is to know about other people.
(2)′ Mary (before her release) does not know everything there is to know about other people (because she *learns* something about them on her release).

Therefore,

(3)′ There are truths about other people (and herself) which escape the physicalist story.

What is immediately to the point is not the kind, manner, or type of knowledge Mary has, but *what* she knows. What she knows beforehand is *ex hypothesi* everything physical there is to know, but is it everything there is to know? That is the crucial question.

NOTES

1. See, e.g., D. H. Mellor, "Materialism and Phenomenal Qualities," *Aristotelian Society Supp. Vol.* 47 (1973): pp. 107–19; and J. W. Cornman, *Materialism and Sensations* (New Haven and London: Yale University Press, 1971).
2. Particularly in discussion, but see, e.g., Keith Campbell; *Metaphysics* (Belmont, CA: Dickenson,1976), p. 67.
3. See, e.g., D. C. Dennett, "Current Issues in the Philosophy of Mind," *American Philosophical Quarterly* 15 (1978): pp. 249–61.

4. Put this, and similar simplifications below, in terms of Land's theory if you prefer. See, e.g., Edwin H. Land, "Experiments in Color Vision," *Scientific American* 200 (May 5, 1959): pp. 84–99.
5. H. G. Wells, *The Country of the Blind and Other Stories* (London, 1911).
6. See, e.g., Keith Campbell, *Body and Mind* (New York: Doubleday Anchor Books, 1970); and Robert Kirk, "Sentience and Behaviour," *Mind* 83 (1974): pp. 43–60.
7. I have presented the argument in an inter-world rather than the more usual intra-world fashion to avoid inessential complications to do with supervenience, causal anomalies, and the like.
8. See, e.g., W. G. Lycan, "A New Lilliputian Argument against Machine Functionalism," *Philosophical Studies* 35 (1979): pp. 279–87; and Don Locke, "Zombies, Schizophrenics and Purely Physical Objects," *Mind* 85 (1976): pp. 97–99.
9. See R. Kirk, "From Physical Explicability to Full-Blooded Materialism," *The Philosophical Quarterly* 29 (1979): pp. 229–37. See also the arguments against the modal intuition in, e.g., Sydney Shoemaker, "Functionalism and Qualia," *Philosophical Studies* 27 (1975): pp. 291–315.
10. *The Philosophical Review,* 83 (1974): pp. 435–50. Two things need to be said about this article. One is that, despite my dissociations to come, I am much indebted to it. The other is that the emphasis changes through the article, and by the end Nagel is objecting not so much to Physicalism as to all extant theories of mind for ignoring points of view, including those that admit (irreducible) qualia.
11. Knowledge *de se* in the terms of David Lewis, "Attitudes De Dicto and De Se," *The Philosophical Review* 88 (1979): pp. 513–43.
12. See Laurence Nemirow's comments on "What is it . . ." in his review of T. Nagel, *Mortal Questions,* in *The Philosophical Review* 89 (1980): pp. 473–77. I am indebted here in particular to a discussion with David Lewis.
13. See my review of K. Campbell, *Body and Mind,* in *Australasian Journal of Philosophy* 50 (1972): pp. 77–80.
14. Cf. Jean Piaget, "The Child's Conception of Physical Causality," reprinted in *The Essential Piaget* (New York: Basic Books, 1977).
15. I am indebted to Robert Pargetter for a number of comments and, despite his dissent, to §IV of Paul E. Meehl, "The Compleat Autocerebroscopist," in *Mind, Matter, and Method,* Paul Feyerabend and Grover Maxwell, ed., (Minneapolis: University of Minnesota Press, 1966).
16. Paul M. Churchland, "Reduction, Qualia, and the Direct Introspection of Brain States," *The Journal of Philosophy,* LXXXII, no. 1 (January 1985): pp. 8–28. Unless otherwise stated, future page references are to this paper.

What Experience Teaches

David Lewis

Experience the Best Teacher

They say that experience is the best teacher, and the classroom is no substitute for Real Life. There's truth to this. If you want to know what some new and different experience is like, you can learn it by going out and really *having* that experience. You can't learn it by being told about the experience, however thorough your lessons may be.

Does this prove much of anything about the metaphysics of mind and the limits of science? I think not.

Example: Skunks and Vegemite

I have smelled skunks, so I know what it's like to smell skunks. But skunks live only in some parts of the world, so you may never have smelled a skunk. If you haven't smelled a skunk, then you don't know what it's like. You never will, unless someday you smell a skunk for yourself. On the other hand, you may have tasted Vegemite, that famous Australian substance; and I never have. So you may know what it's like to taste Vegemite. I don't, and unless I taste Vegemite (what, and spoil a good

Reprinted from *Proceedings of the Russellian Society* (University of Sydney, 1988), with permission of the author's estate and the publisher.

example!), I never will. It won't help at all to take lessons on the chemical composition of skunk scent or Vegemite, the physiology of the nostrils or the taste-buds, and the neurophysiology of the sensory nerves and the brain.

Example: The Captive Scientist[1]

Mary, a brilliant scientist, has lived from birth in a cell where everything is black or white. (Even she herself is painted all over.) She views the world on black-and-white television. By television she reads books, she joins in discussion, she watches the results of experiments done under her direction. In this way she becomes the world's leading expert on color and color vision and the brain states produced by exposure to colors. But she doesn't know what it's like to see color. And she never will, unless she escapes from her cell.

Example: The Bat[2]

The bat is an alien creature, with a sonar sense quite unlike any sense of ours. We can never have the experiences of a bat; because we could not become bat-like enough to have those experiences and still be ourselves. We will never know what it's like to be a bat. Not even if we come to know all the facts there are about the bat's behavior and behavioral dispositions, about the bat's physical structure and processes, about the bat's functional organization. Not even if we come to know all the same sort of physical facts about all the other bats, or about other creatures, or about ourselves. Not even if we come to possess all physical facts whatever. Not even if we become able to recognize all the mathematical and logical implications of all these facts, no matter how complicated and how far beyond the reach of finite deduction.

Experience is the best teacher, in this sense: having an experience is the best way or perhaps the only way, of coming to know what that experience is like. No amount of scientific information about the stimuli that produce that experience and the process that goes on in you when you have that experience will enable you to know what it's like to have the experience.

. . . But Not Necessarily

Having an experience is surely one good way, and surely the only practical way, of coming to know what that experience is like. Can we say, flatly, that it is the only *possible* way? Probably not. There is a change that takes place in you when you have the experience and thereby come to know what it's like. Perhaps the exact same change could in principle be produced in you by precise neurosurgery, very far beyond the limits of present-day technique. Or it could possibly be produced in you by magic. If we ignore the laws of nature, which are after all contingent, then there is no necessary connection between cause and effect: anything could cause anything. For instance, the casting of a spell could do to you exactly what your first smell of skunk would do. We might quibble about whether a state produced in this artificial fashion would deserve the *name* 'knowing what it's like to smell a skunk,' but we can imagine that so far as what goes on within you is concerned, it would differ not at all.[3]

Just as we can imagine that a spell might produce the same change as a smell, so likewise we can imagine that science lessons might cause that same change. Even that is possible, in the broadest sense of the word. If we ignored all we know about how the world really works, we could not say what might happen to someone if he were taught about the chemistry of scent and the physiology of the nose. There might have been a causal mechanism that transforms science lessons into whatever it is that experience gives us. But there isn't. It is not an absolutely necessary truth that experience is the best teacher about what a new experience is like. It's a contingent truth. But we have good reason to think it's true.

We have good reason to think that something of this kind is true, anyway, but less reason to be sure exactly what. Maybe some way of giving the lessons that hasn't yet been invented, and some way of taking them in that hasn't yet been practiced, could give us a big surprise. Consider sight-reading: a trained musician can read the score and know what it would be like to hear the music. If I'd never heard that some people can sight-read, I would never have thought it humanly possible. Of course the moral is that new music isn't altogether new—the big new experience is a rearrangement of lots of little old experiences. It just might turn out the same for new smells and tastes vis-à-vis old ones; or even for color vision vis-à-vis black and white;[4] or even for sonar sense experience vis-à-vis the sort we enjoy. The thing we can say with some confidence is that we have no faculty for knowing on the basis of mere science lessons what some *new enough* experience would be like. But how new is 'new enough'?—There, we just might be in for surprises.

Three Ways to Miss the Point

The First Way

A literalist might see the phrase 'know what it's like' and take that to mean: 'know what it resembles.' Then he might ask: what's so hard about that? Why can't you just be told which experiences resemble one another? You needn't have had the experiences—all you need, to be taught your lessons, is some way of referring to them. You could be told: the smell of skunk somewhat resembles the smell of burning rubber. I have been told: the taste of Vegemite somewhat resembles that of Marmite. Black and-white Mary might know more than most of us about the resemblances among color-experiences. She might know which ones are spontaneously called 'similar' by subjects who have them; which gradual changes from one to another tend to escape notice; which ones get conflated with which in memory; which ones involve roughly the same neurons firing in similar rhythms; and so forth. We could even know what the bat's sonar experiences resemble just by knowing that they do not at all resemble any experiences of humans, but do resemble—as it might be—certain experiences that occur in certain fish. This misses the point. Pace the literalist, 'know what it's like' does not mean 'know what it resembles.' The most that's true is that knowing what it resembles *may* help you to know what it's like. If you are taught that experience A resembles B and C closely, D less, E not at all, that will help you know what A is like—*if* you know already what B and C and D and E are like. Otherwise, it helps you not at all. I don't know any better what it's like to taste Vegemite when I'm told that it tastes like Marmite, because I don't know what Marmite tastes like either. (Nor do I know any better what Marmite tastes like for being told it tastes like Vegemite.) Maybe Mary knows enough to triangulate each color experience exactly in a network of resemblances, or in many networks of resemblance in different respects, while never knowing what any node of any network is like. Maybe we could do the same for bat experiences. But no amount of information about resemblances, just by itself, does anything to help us know what an experience is like.

The Second Way

In so far as I don't know what it would be like to drive a steam locomotive fast on a cold, stormy night, part of my problem is just that I don't know what experiences I would have. The fire-box puts out a lot of heat, especially when the fireman opens the door to throw on more coal; on the other hand, the cab is drafty and gives poor protection from the weather. Would I be too hot or too cold? Or both by turns? Or would it be chilled face and scorched legs? If I knew the answers to such questions, I'd know much better what it would be like to drive the locomotive. So maybe 'know what it's like' just means 'know what experiences one has.' Then again: what's the problem? Why can't you just be told what experiences you would have if, say, you tasted Vegemite? Again, you needn't have had the experiences—all you need, to be taught your lessons, is some way of referring to them. We have ways to refer to experiences we haven't had. We can refer to them in terms of their causes: the experience one has upon tasting Vegemite, the experience one has upon tasting a substance of such-and-such chemical composition. Or we can refer to them in terms of their effects: the experience that just caused Fred to say 'Yeeuch!' Or we can refer to them in terms of the physical states of the nervous system that mediate between those causes and effects: the experience one has when one's nerves are firing in such-and-such pattern. (According to some materialists, I myself for one, this means the experience which is identical with such-and-such firing pattern. According to other materialists it means the experience which is realized by such-and-such firing pattern. According to many dualists, it means the experience which is merely the lawful companion of such-and-such firing pattern. But whichever it is, we get a way of referring to the experience.) Black-and-white Mary is in a position to refer to color-experiences in all these ways. Therefore you should have no problem in telling her exactly what experiences one has upon seeing the colors. Or rather, your only problem is that you'd be telling her what she knows very well already! In general, to know what is the X is to know that the X is the Y, where it's not too obvious that the X is the Y. (Just knowing that the X is the X won't do, of course, because it is too obvious.) If Mary knows that the experience of seeing green is the experience associated with such-and-such pattern of nerve firings, then she knows the right sort of unobvious identity. So she knows what experience one has upon seeing green.

(Sometimes it's suggested that you need a 'rigid designator': you know what is the X by knowing that the X is the Y only if 'the Y'

is a term whose referent does not depend on any contingent matter of fact. In the first place, this suggestion is false. You can know who is the man on the balcony by knowing that the man on the balcony is the Prime Minister even if neither 'the Prime Minister' nor any other phrase available to you rigidly designates the man who is, in fact, the Prime Minister. In the second place, according to one version of Materialism [the one I accept] a description of the form 'the state of having nerves firing in such-and-such a pattern' *is* a rigid designator, and what it designates is in fact an experience; and according to another version of Materialism, a description of the form 'having some or other state which occupies so-and-so functional role' is a rigid designator of an experience. So even if the false suggestion were granted, still it hasn't been shown, without begging the question against Materialism, that Mary could not know what experience one has upon seeing red.)

Since Mary *does* know what experiences she would have if she saw the colors, but she *doesn't* know what it would be like to see the colors, we'd better conclude that 'know what it's like' does not after all mean 'know what experiences one has.' The locomotive example was misleading. Yes, by learning what experiences the driver would have, I can know what driving the locomotive would be like; but only because I already know what those experiences are like. (It matters that I know what they're like under the appropriate descriptions—as it might be, the description 'chilled face and scorched legs.' This is something we'll return to later.) Mary may know as well as I do that when the driver leans out into the storm to watch the signals, he will have the experience of seeing sometimes green lights and sometimes red. She knows better than I what experiences he has when signals come into view. She can give many more unobviously equivalent descriptions of those experiences than I can. But knowing what color-experiences the driver has won't help Mary to know what his job is like. It will help me.

The Third Way

Until Mary sees green, here is one thing she will never know: she will never know that she is seeing green. The reason why is just that until she sees green, it will never be true that she is seeing green. Some knowledge is irreducibly egocentric, or *de se*.[5] It is not just knowledge about what goes on in the world; it is knowledge of who and when in the world one is. Knowledge of what goes on in the world will be true alike for all who live in that world; whereas egocentric knowledge may be true for one and false for another, or true for one at one time and false for the same one at another time. Maybe Mary knows in advance, as she plots her escape, that 9 A.M. on the 13th of May, 1997, is the moment when someone previously confined in a black-and-white cell sees color for the first time. But until that moment comes, she will never know that she herself is then seeing color—because she isn't. What isn't true isn't knowledge. This goes as much for egocentric knowledge as for the rest. So only those of whom an egocentric proposition is true can know it, and only at times when it is true of them can they know it. That one is then seeing color is an egocentric proposition. So we've found a proposition which Mary can never know until she sees color—which, as it happens, is the very moment when she will first know what it's like to see color! Have we discovered the reason why experience is the best teacher? And not contingently after all, but as a necessary consequence of the logic of egocentric knowledge?

No; we have two separate phenomena here, and only some bewitchment about the 'first-person perspective' could make us miss the difference. In the first place, Mary will probably go on knowing what it's like to see green after she stops knowing the egocentric proposition that she's then seeing green. Since what isn't true isn't known she must stop knowing that proposition the moment she stops seeing green. (Does that only mean that we should have taken a different egocentric proposition: that one *has* seen green? No; for in that case Mary could go on knowing the proposition even after she forgets what it's like to see green, as might happen if she were soon recaptured.) In the second place, Mary might come to know what it's like to see green even if she didn't know the egocentric proposition. She might not have known in advance that her escape route would take her across a green meadow, and it might take her a little while to recognize grass by its shape. So at first she might know only that she was seeing some colors or other, and thereby finding out what some color-experiences or other were like, without being able to put a name either to the colors or to the experiences. She would then know what it was like to see green, though

not under that description, indeed not under any description more useful than 'the color-experience I'm having now'; but she would not know the egocentric proposition that she is then seeing green, since she wouldn't know which color she was seeing. In the third place, the gaining of egocentric knowledge may have prerequisites that have nothing to do with experience. Just as Mary can't know she's seeing green until she *does* see green, she can't know she's turning 50 until she *does* turn 50. But—I hope!—turning 50 does not involve some special experience. In short, though indeed one can gain egocentric knowledge that one is in some situation only when one is in it, that is not the same as finding out what an experience is like only when one has that experience.

We've just rejected two suggestions that don't work separately, and we may note that they don't work any better when put together. One knows what is the X by knowing that the X is the Y, where the identity is not too obvious; and 'the Y' might be an egocentric description. So knowledge that the X is the Y might be irreducibly egocentric knowledge, therefore knowledge that cannot be had until it is true of one that the X is the Y. So one way of knowing what is the X will remain unavailable until it comes true of one that the X is the Y. One way that I could gain an unobvious identity concerning the taste of Vegemite would be for it to come true that the taste of Vegemite was the taste I was having at that very moment—and that would come true at the very moment I tasted Vegemite and found out what it was like! Is this why experience is the best teacher?—No; cases of gaining an unobvious egocentric identity are a dime a dozen, and most of them do not result in finding out what an experience is like. Suppose I plan ahead that I will finally break down and taste Vegemite next Thursday noon. Then on Wednesday noon, if I watch the clock, I first gain the unobvious egocentric knowledge that the taste of Vegemite is the taste I shall be having in exactly 24 hours, and thereby I have a new way of knowing what is the taste of Vegemite. But on Wednesday noon I don't yet know what it's like. Another example: from time to time I find myself next to a Vegemite-taster. On those occasions, and only those, I know what is the taste of Vegemite by knowing that it is the taste being had by the person next to me. But on no such occasion has it ever yet happened that I knew what it was like to taste Vegemite.

The Hypothesis of Phenomenal Information

No amount of the physical information that black-and-white Mary gathers could help her know what it was like to see colors; no amount of the physical information that we might gather about bats could help us know what it's like to have their experiences; and likewise in other cases. There is a natural and tempting explanation of why physical information does not help. That is the hypothesis that besides physical information there is an irreducibly different kind of information to be had: *phenomenal information.* The two are independent. Two possible cases might be exactly alike physically, yet differ phenomenally. When we get physical information we narrow down the physical possibilities, and perhaps we narrow them down all the way to one, but we leave open a range of phenomenal possibilities. When we have an experience, on the other hand, we acquire phenomenal information; possibilities previously open are eliminated; and that is what it is to learn what the experience is like.

(Analogy. Suppose the question concerned the location of a point within a certain region of the x-y plane. We might be told that its x-coordinate lies in certain intervals, and outside certain others. We might even get enough of this information to fix the x-coordinate exactly. But no amount of x-information would tell us anything about the y-coordinate; any amount of x-information leaves open all the y-possibilities. But when at last we make a y-measurement, we acquire a new kind of information; possibilities previously open are eliminated; and that is how we learn where the point is in the y-direction.)

What might the subject matter of phenomenal information be? *If* the Hypothesis of Phenomenal Information is true, then you have an easy answer: it is information about experience. More specifically, it is information about a certain part or aspect or feature of experience. But if the Hypothesis is false, then there is still experience (complete with all its parts and aspects and features) and yet no information about experience is phenomenal information. So it cannot be said in a neutral way, without presupposing the Hypothesis, that information about experience is phenomenal information. For if the Hypothesis is false and Materialism is true, it may be that all the information there is about experience is physical information, and can very well be presented in lessons for the inexperienced.

It makes no difference to put some fashionable new phrase in place of 'experience.' If instead of 'experience' you say 'raw feel' (or just 'feeling'), or 'way it feels,' or 'what it's like,' then I submit that you mean nothing different. Is there anything it's like to be this robot? Does this robot have experiences?—I can tell no difference between the new question and the old. Does sunburn feel the same way to you that it does to me? Do we have the same raw feel? Do we have the same experience when sunburned?—Again, same question. 'Know the feeling,' 'know what it's like'—interchangeable. (Except that the former may hint at an alternative to the Hypothesis of Phenomenal Information.) So if the friend of phenomenal information says that its subject matter is raw feels, or ways to feel, or what it's like, then I respond just as I do if he says that the subject matter is experience. Maybe so, *if* the Hypothesis of Phenomenal Information is true; but if the Hypothesis is false and Materialism is true, nevertheless there is still information about raw feels, ways to feel or what it's like; but in that case it is physical information and can be conveyed in lessons.

We might get a candidate for the subject matter of phenomenal information that is not just experience renamed, but is still tendentious. For instance, we might be told that phenomenal information concerns the intrinsic character of experience. A friend of phenomenal information might indeed believe that it reveals certain special, non-physical intrinsic properties of experience. He might even believe that it reveals the existence of some special non-physical thing or process, *all* of whose intrinsic properties are non-physical. But he is by no means alone in saying that experience has an intrinsic character. Plenty of us materialists say so too. We say that a certain color-experience is whatever state occupies a certain functional role. So if the occupant of that role (universally, or in the case of humans, or in the case of certain humans) is a certain pattern of neural firing, then that pattern of firing *is* the experience (in the case in question). Therefore the intrinsic character of the experience is the intrinsic character of the firing pattern. For instance, a frequency of firing is part of the intrinsic character of the experience. If we materialists are right about what experience is, then black-and-white Mary knows all about the intrinsic character of color-experience; whereas most people who know what color-experience is like remain totally ignorant about its intrinsic character.[6]

To say that phenomenal information concerns 'qualia' would be tendentious in much the same way. For how was this notion introduced? Often thus. We are told to imagine someone who, when he sees red things, has just the sort of experiences that we have when we see green things, and vice versa; and we are told to call this a case of 'inverted qualia.' And then we are told to imagine someone queerer still, who sees red and responds to it appropriately, and indeed has entirely the same functional organization of inner states as we do and yet has no experiences at all; and we are told to call this a case of 'absent qualia.' Now a friend of phenomenal information might well think that these deficiencies have something to do with the non-physical subject matter of phenomenal information. But others can understand them otherwise. Some materialists will reject the cases outright, but others, and I for one, will make sense of them as best we can. Maybe the point is that the states that occupy the roles of experiences, and therefore *are* the experiences, in normal people are inverted or absent in victims of inverted or absent qualia. (This presupposes, what might be false, that most people are enough alike). Experience of red—the state that occupies that role in normal people—occurs also in the victim of 'inverted qualia,' but in him it occupies the role of experience of green; whereas the state that occupies in him the role of experience of red is the state that occupies in normal people the role of experience of green. Experience of red and of green—that is, the occupants of those roles for normal people—do not occur at all in the victim of 'absent qualia'; the occupants of those roles for him are states that don't occur at all in the normal. Thus we make good sense of inverted and absent qualia; but in such a way that 'qualia' is just the word for role-occupying states taken *per se* rather than *qua* occupants of roles. Qualia, so understood, could not be the subject matter of phenomenal information. Mary knows all about them. We who have them mostly don't.[7]

It is best to rest content with an unhelpful name and a *via negativa.* Stipulate that 'the phenomenal aspect of the world' is to name whatever is the subject matter of phenomenal information, if there is any such thing; the phenomenal aspect, if such there be, is that which we can become informed about by having new experiences but never by taking lessons. Having said this, it will be safe to say that information about the phenomenal aspect of the world can only be phenomenal information.

But all we really know, after thus closing the circle, is that phenomenal information is supposed to reveal the presence of some sort of non-physical things or processes within experience, or else it is supposed to reveal that certain physical things or processes within experience have some sort of nonphysical properties.

The Knowledge Argument

If we invoke the Hypothesis of Phenomenal Information to explain why no amount of physical information suffices to teach us what a new experience is like, then we have a powerful argument to refute any materialist theory of the mind. Frank Jackson (see note 1) calls it the 'Knowledge Argument.' Arguments against one materialist theory or another are never very conclusive. It is always possible to adjust the details. But the Knowledge Argument, if it worked, would directly refute the bare minimum that is common to *all* materialist theories.

It goes as follows. First in a simplified form; afterward we'll do it properly. Minimal Materialism is a supervenience thesis: no difference without physical difference. That is: any two possibilities that are just alike physically are just alike *simpliciter.* If two possibilities are just alike physically, then no physical information can eliminate one but not both of them. If two possibilities are just alike *simpliciter* (if that is possible) then no information whatsoever can eliminate one but not both of them. So if there is a kind of information—namely, phenomenal information—that can eliminate possibilities that any amount of physical information leaves open, then there must be possibilities that are just alike physically, but not just alike *simpliciter.* That is just what minimal Materialism denies.

(Analogy. If two possible locations in our region agree in their x-coordinate, then no amount of x-information can eliminate one but not both. If, *per impossible,* two possible locations agreed in all their coordinates, then no information whatsoever could eliminate one but not both. So if there is a kind of information—namely, y-information—that can eliminate locations that any amount of x-information leaves open, then there must be locations in the region that agree in their x-coordinate but not in all their coordinates.)

Now to remove the simplification. What we saw so far was the Knowledge Argument against Materialism taken as a necessary truth, applying unrestrictedly to all possible worlds. But we materialists usually think that Materialism is a contingent truth. We grant that there are spooky possible worlds where Materialism is false, but we insist that our actual world isn't one of them. If so, then there might after all be two possibilities that are alike physically but not alike *simpliciter;* but one or both of the two would have to be possibilities where Materialism was false. Spooky worlds could differ with respect to their spooks without differing physically. Our minimal Materialism must be a *restricted* supervenience thesis: within a certain class of worlds, which includes our actual world, there is no difference without physical difference. Within that class, any two possibilities just alike physically are just alike *simpliciter.* But what delineates the relevant class? (It is trivial that our world belongs to *some* class wherein there is no difference without physical difference. That will be so however spooky our world may be. The unit class of our world is one such class, for instance. And so is any class that contains our world, and contains no two physical duplicates.) I think the relevant class should consist of the worlds that have nothing wholly alien to this world. The inhabitants of such a non-alien world could be made from the inhabitants of ours, so to speak, by a process of division and recombination. That will make no wholly different kinds of things, and no wholly different fundamental properties of things.[8] Our restricted materialist supervenience thesis should go as follows: throughout the non-alien worlds, there is no difference without physical difference.

If the Hypothesis of Phenomenal Information be granted, then the Knowledge Argument refutes this restricted supervenience nearly as decisively as it refutes the unrestricted version. Consider a possibility that is eliminated by phenomenal information, but not by any amount of physical information. There are two cases. Maybe this possibility has nothing that is alien to our world. In that case the argument goes as before: actuality and the eliminated possibility are just alike physically, they are not just alike *simpliciter;* furthermore, both of them fall within the restriction to non-alien worlds, so we have a counterexample even to restricted supervenience. Or maybe instead the eliminated possibility does have something X which is alien to this world—an alien kind of thing, or maybe an alien fundamental property of non-alien things. Then the phenomenal information gained by

having a new experience has revealed something negative: at least in part, it is the information that X is *not* present. How can that be? If there is such a thing as phenomenal information, presumably what it reveals is positive: the presence of something hitherto unknown. Not, of course, something alien from actuality itself; but something alien from actuality as it is inadequately represented by the inexperienced and by the materialists. If Mary learns something when she finds out what it's like to see the colors, presumably she learns that there's *more* to the world than she knew before—not *less*. It's easy to think that phenomenal information might eliminate possibilities that are impoverished by comparison with actuality, but that would make a counterexample to the restricted supervenience thesis. To eliminate possibilities without making a counterexample, phenomenal information would have to eliminate possibilities less impoverished than actuality. And how can phenomenal information do that? Compare ordinary perceptual information. Maybe Jean-Paul can just *see* that Pierre is absent from the café, at least if it's a small café. But how can he just see that Pierre is absent from Paris, let alone from the whole of actuality?

(Is there a third case? What if the eliminated possibility is in one respect richer than actuality, in another respect poorer? Suppose the eliminated possibility has X, which is alien from actuality, but also it lacks Y. Then phenomenal information might eliminate it by revealing the actual presence of Y, without having to reveal the actual absence of X—But then I say there ought to be a third possibility, one with neither X nor Y, poorer and in no respect richer than actuality, and again without any physical difference from actuality. For why should taking away X automatically restore Y? Why can't they vary independently?[9] But this third possibility differs *simpliciter* from actuality without differing physically. Further, it has nothing alien from actuality. So we regain a counterexample to the restricted supervenience thesis.)

The Knowledge Argument works. There is no way to grant the Hypothesis of Phenomenal Information and still uphold Materialism. Therefore I deny the Hypothesis. I cannot refute it outright. But later I shall argue, first, that it is more peculiar, and therefore less tempting, that it may at first seem; and, second, that we are not forced to accept it, since an alternative hypothesis does justice to the way experience best teaches us what it's like.

Three More Ways to Miss the Point

The Hypothesis of Phenomenal Information characterizes information in terms of eliminated possibilities. But there are other conceptions of 'information.' Therefore the Hypothesis has look-alikes: hypotheses which say that experience produces 'information' which could not be gained otherwise, but do not characterize this 'information' in terms of eliminated possibilities. These look-alikes do not work as premises for the Knowledge Argument. They do not say that phenomenal information eliminates possibilities that differ, but do not differ physically, from uneliminated possibilities. The look-alike hypotheses of phenomenal 'information' are consistent with Materialism, and may very well be true. But they don't make the Knowledge Argument go away. Whatever harmless look-alikes may or may not be true, and whatever conception may or may not deserve the name 'information,' the only way to save Materialism is fix our attention squarely on the genuine Hypothesis of Phenomenal Information, and deny it. To avert our eyes, and attend to something else, is no substitute for that denial.

Might a look-alike help at least to this extent: by giving us something true that well might have been confused with the genuine Hypothesis, thereby explaining how we might have believed the Hypothesis although it was false? I think not. Each of the look-alikes turns out to imply not only that experience can give us 'information' that no amount of lessons can give, but also that lessons in Russian can give us 'information' that no amount of lessons in English can give (and vice versa). I doubt that any friend of phenomenal information ever thought that the special role of experience in teaching what it's like was on a par with the special role of Russian! I will have to say before I'm done that phenomenal information is an illusion, but I think I must look elsewhere for a credible hypothesis about what sort of illusion it might be.

The Fourth Way

If a hidden camera takes photographs of a room, the film ends up bearing traces of what went on in the room. The traces are distinctive: that is, the details of the traces depend on the details of what went on, and if what went on had been different in any of many ways, the traces would have been correspondingly different. So we can

say that the traces bear information, and that he who has the film has the information. That might be said because the traces, plus the way they depend on what went on, suffice to eliminate possibilities; but instead we might say 'information' and just mean 'distinctive traces.' If so, it's certainly true that new experience imparts 'information' unlike any that can be gained from lessons. Experience and lessons leave different kinds of traces. That is so whether or not the experience eliminates possibilities that the lessons leave open. It is equally true, of course, that lessons in Russian leave traces unlike any that are left by lessons in English, regardless of whether the lessons cover the same ground and eliminate the same possibilities.

The Fifth Way

When we speak of transmission of 'information,' we often mean transmission of text. Repositories of 'information,' such as libraries, are storehouses of text. Whether the text is empty verbiage or highly informative is beside the point. Maybe we too contain information by being storehouses of text. Maybe there is a language of thought, and maybe the way we believe things is to store sentences of this language in some special way, or in some special part of our brains. In that case, we could say that storing away a new sentence was storing away a new piece of 'information,' whether or not that new piece eliminated any possibilities not already eliminated by the sentences stored previously. Maybe, also, the language of thought is not fixed once and for all, but can gain new words. Maybe, for instance, it borrows words from public language. And maybe, when one has a new experience, that causes one's language of thought to gain a new word which denotes that experience—a word which could not have been added to the language by any other means. If all this is so, then when Mary sees colors, her language of thought gains new words, allowing her to store away new sentences and thereby gain 'information.' All this about the language of thought, the storing of sentences, and the gaining of words is speculation. But it is plausible speculation, even if no longer the only game in town. If it is all true, then we have another look-alike hypothesis of phenomenal 'information.' When Mary gains new words and stores new sentences, that is 'information' that she never had before, regardless of whether it eliminates any possibilities that she had not eliminated already.

But again, the special role of experience turns out to be on a par with the special role of Russian. If the language of thought picks up new words by borrowing from public language, then lessons in Russian add new words, and result in the storing of new sentences, and thereby impart 'information' that never could have been had from lessons in English. (You might say that the new Russian words are mere synonyms of old words, or at least old phrases, that were there already; and synonyms don't count. But no reason has been given why the new inner words created by experience may not also be synonyms of old phrases, perhaps of long descriptions in the language of neurophysiology.)

The Sixth Way

A philosopher who is skeptical about possibility, as so many are, may wish to replace possibilities themselves with linguistic ersatz possibilities: maximal consistent sets of sentences. And he may be content to take 'consistent' in a narrowly logical sense, so that a set with 'Fred is married' and 'Fred is a bachelor' may count as consistent, and only an overt contradiction like 'Fred is married' and 'Fred is not married' will be ruled out.[10] The ersatz possibilities might also be taken as sets of sentences of the language of thought, if the philosopher believes in it. Then if someone's language of thought gains new words, whether as a result of new experience or as a result of being taught in Russian, the ersatz possibilities become richer and more numerous. The sets of sentences that were maximal before are no longer maximal after new words are added. So when Mary sees colors and her language of thought gains new words, there are new ersatz possibilities; and she can straightway eliminate some of them. Suppose she knows beforehand that she is about to see green, and that the experience of seeing green is associated with neural firing pattern F. So when she sees green and gains the new word G for her experience, then straightway there are new, enriched ersatz possibilities with sentences saying that she has G without F, and straightway she knows enough to eliminate these ersatz possibilities. (Even if she does not know beforehand what she is about to see, straightway she can eliminate at least those of her new-found ersatz possibilities with sentences denying that she then has G.) Just as we can characterize information in terms of elimination of possibilities, so we can characterize ersatz 'information' in

terms of elimination of ersatz 'possibilities.' So here we have the closest look-alike hypothesis of all, provided that language-of-thoughtism is true. But we still do not have the genuine Hypothesis of Phenomenal Information, since the eliminated ersatz possibility of G without F may not have been a genuine possibility at all. It may have been like the ersatz possibility of married bachelors.

Curiouser and Curiouser

The Hypothesis of Phenomenal Information is more peculiar than it may at first seem. For one thing, because it is opposed to more than just Materialism. Some of you may have welcomed the Knowledge Argument because you thought all along that physical information was inadequate to explain the phenomena of mind. You may have been convinced all along that the mind could do things that no physical system could do: bend spoons, invent new jokes, demonstrate the consistency of arithmetic, reduce the wave packet, or what have you. You may have been convinced that the full causal story of how the deeds of mind are accomplished involves the causal interactions not only of material bodies but also of astral bodies; not only the vibrations of the electromagnetic field but also the good or bad vibes of the psionic field; not only protoplasm but ectoplasm. I doubt it, but never mind. It's irrelevant to our topic. The Knowledge Argument is targeted against you no less than it is against Materialism itself.

Let *parapsychology* be the science of all the non-physical things, properties, causal processes, laws of nature, and so forth that may be required to explain the things we do. Let us suppose that we learn ever so much parapsychology. It will make no difference. Black-and-white Mary may study all the parapsychology as well as all the psychophysics of color vision, but she still won't know what it's like. Lessons on the aura of Vegemite will do no more for us than lessons on its chemical composition. And so it goes. Our intuitive starting point wasn't just that *physics* lessons couldn't help the inexperienced to know what it's like. It was that *lessons* couldn't help. If there is such a thing as phenomenal information, it isn't just independent of physical information. It's independent of every sort of information that could be served up in lessons for the inexperienced. For it is supposed to eliminate possibilities that any amount of lessons leave open. Therefore phenomenal information is not just parapsychological information, if such there be. It's something very much stranger.

The genuine Hypothesis of Phenomenal Information, as distinguished from its look-alikes, treats information in terms of the elimination of possibilities. When we lack information, several alternative possibilities are open, when we get the information some of the alternatives are excluded. But a second peculiar thing about phenomenal information is that it resists this treatment. (So does logical or mathematical 'information.' However, phenomenal information cannot be logical or mathematical, because lessons in logic and mathematics no more teach us what a new experience is like than lessons in physics or parapsychology do.) When someone doesn't know what it's like to have an experience, where are the alternative open possibilities? I cannot present to myself in thought a range of alternative possibilities about what it might be like to taste Vegemite. That is because I cannot imagine either what it *is* like to taste Vegemite, or any alternative way that it *might* be like but in fact isn't. (I could perfectly well imagine that Vegemite tastes just like peanut butter, or something else familiar to me, but let's suppose I've been told authoritatively that this isn't so.) I can't even pose the question that phenomenal information is supposed to answer: is it this way or that? It seems that the alternative possibilities must be unthinkable beforehand; and afterward too, except for the one that turns out to be actualized. I don't say there's anything altogether impossible about a range of unthinkable alternatives; only something peculiar. But it's peculiar enough to suggest that we may somehow have gone astray.

From Phenomenal to Epiphenomenal

A third peculiar thing about phenomenal information is that it is strangely isolated from all other sorts of information; and this is so regardless of whether the mind works on physical or parapsychological principles. The phenomenal aspect of the world has nothing to do with explaining why people seemingly talk about the phenomenal aspect of the world. For instance, it plays no part in explaining the movements of the pens of philosophers writing treatises about phenomenal information and the way experience has provided them with it.

When Mary gets out of her black-and-white cell, her jaw drops. She says 'At last! So this is what it's like to see colors!' Afterward she does things she couldn't do before, such as recognizing a new sample of the first color she ever saw. She may also do other things she didn't do before: unfortunate things, like writing about phenomenal information and the poverty of Materialism. One might think she said what she said and did what she did because she came to know what it's like to see colors. Not so, if the Hypothesis of Phenomenal Information is right. For suppose the phenomenal aspect of the world had been otherwise, so that she gained different phenomenal information. Or suppose the phenomenal aspect of the world had been absent altogether, as we materialists think it is. Would that have made the slightest difference to what she did or said then or later? I think not. Making a difference to what she does or says means, at least in part, making a difference to the motions of the particles of which she is composed. (Or better: making a difference to the spatiotemporal shape of the wave-function of those particles. But let that pass.) For how could she do or say anything different, if none of her particles moved any differently? But if something non-physical sometimes makes a difference to the motions of physical particles, then physics as we know it is wrong. Not just silent, not just incomplete—wrong. Either the particles are caused to change their motion without benefit of any force, or else there is some extra force that works very differently from the usual four. To believe in the phenomenal aspect of the world, but deny that it is epiphenomenal, is to bet against the truth of physics. Given the success of physics hitherto, and even with due allowance for the foundational ailments of quantum mechanics, such betting is rash! A friend of the phenomenal aspect would be safer to join Jackson in defense of *epiphenomenal* qualia.

But there is more to the case than just an empirical bet in favor of physics. Suppose there is a phenomenal aspect of the world, and suppose it does make some difference to the motions of Mary's jaw or the noises out of her mouth. Then we can describe the phenomenal aspect, if we know enough, in terms of its physical effects. It is that on which physical phenomena depend in such-and-such way. This descriptive handle will enable us to give lessons on it to the inexperienced. But in so far as we can give lessons on it, what we have is just parapsychology. That whereof we cannot learn except by having the experience still eludes us. I do not argue that *everything* about the alleged distinctive subject matter of phenomenal information must be epiphenomenal. Part of it may be parapsychological instead. But I insist that *some* aspect of it must be epiphenomenal.

Suppose that the Hypothesis of Phenomenal Information is true and suppose that V_1 and V_2 are all of the maximally specific phenomenal possibilities concerning what it's like to taste Vegemite; anyone who tastes Vegemite will find out which one obtains, and no one else can. And suppose that P_1 and P_2 are all the maximally specific physical possibilities. (Of course we really need far more than two Ps, and maybe a friend of phenomenal information would want more than two Vs, but absurdly small numbers will do for an example.) Then we have four alternative hypotheses about the causal independence or dependence of the Ps on the Vs. Each one can be expressed as a pair of counterfactual conditionals. Two hypotheses are patterns of dependence.

K_1: if V_1 then P_1, if V_2 then P_2
K_2: if V_1 then P_2, if V_2 then P_1

The other two are patterns of independence.

K_3: if V_1 then P_1, if V_2 then P_1
K_4: if V_1 then P_2, if V_2 then P_2

These dependency hypotheses are, I take it, contingent propositions. They are made true, if they are, by some contingent feature of the world, though it's indeed a vexed question what sort of feature it is.[11] Now we have eight joint possibilities.

$K_1V_1P_1$	$K_3V_1P_1$	$K_3V_2P_1$	$K_2V_2P_1$
$K_2V_1P_2$	$K_4V_1P_2$	$K_4V_2P_2$	$K_1V_2P_2$

Between the four on the top row and the four on the bottom row, there is the physical difference between P_1 and P_2. Between the four on the left and the four on the right, there is the phenomenal difference between V_1 and V_2. And between the four on the edges and the four in the middle there is a parapsychological difference. It is the difference between dependence and independence of the physical on the phenomenal; between efficacy and epiphenomenalism, so far as this one example is concerned. There's nothing ineffable about that. Whether or not you've tasted Vegemite, and whether or not you can conceive of the alleged difference between V_1 and V_2, you can still be told whether the physical difference between P_1 and P_2 does or doesn't depend on some part of the phenomenal aspect of the world.

Lessons can teach the inexperienced which parapsychological possibility obtains, dependence or independence. Let it be dependence: we have either K_1 or K_2. For if we had independence, then already we would have found our epiphenomenal difference: namely, the difference between V_1 and V_2. And lessons can teach the inexperienced which of the two physical possibilities obtains. Without loss of generality let it be P_1. Now two of our original eight joint possibilities remain open: $K_1V_1P_1$ and $K_2V_2P_1$. The difference between those is not at all physical, and not at all parapsychological: it's P_1, and its dependence, in both cases. The difference is entirely phenomenal. And also it is entirely epiphenomenal. Nothing physical, and nothing parapsychological, depends on the difference between $K_1V_1P_1$ and $K_2V_2P_1$. We have the same sort of pattern of dependence either way; it's just that the phenomenal possibilities have been swapped. Whether it's independence or whether it's dependence, therefore, we have found an epiphenomenal part of the phenomenal aspect of the world. It is the residue left behind when we remove the parapsychological part.

Suppose that someday I taste Vegemite, and hold forth about how I know at last what it's like. The sound of my holding forth is a physical effect, part of the realized physical possibility P_1. This physical effect is exactly the same whether it's part of the joint possibility $K_1V_1P_1$ or part of its alternative $K_2V_2P_1$. It may be caused by V_1 in accordance with K_1, or it may instead be caused by V_2 in accordance with K_2, but it's the same either way. So it does not occur because we have K_1V_1 rather than K_2V_2, or vice versa. The alleged difference between these two possibilities does nothing to explain the alleged physical manifestation of my finding out which one of them is realized. It is in that way that the difference is epiphenomenal. That makes it very queer, and repugnant to good sense.

The Ability Hypothesis

So the Hypothesis of Phenomenal Information turns out to be very peculiar indeed. It would be nice, and not only for materialists, if we could reject it. For materialists, it is essential to reject it. And we can. There is an alternative hypothesis about what it is to learn what an experience is like: the *Ability Hypothesis*. Laurence Nemirow summarizes it thus:

> some modes of understanding consist, not in the grasping of facts, but in the acquisition of abilities. . . . As for understanding an experience, we may construe that as an ability to place oneself, at will, in a state representative of the experience. I understand the experience of seeing red if I can at will visualize red. Now it is perfectly clear why there must be a special connection between the ability to place oneself in a state representative of a given experience and the point of view of experiencer: exercising the ability just *is* what we call 'adopting the point of view of experiencer.' . . . We can, then, come to terms with the subjectivity of our understanding of experience without positing subjective facts as the objects of our understanding. This account explains, incidentally, the linguistic incommunicability of our subjective understanding of experience (a phenomenon which might seem to support the hypothesis of subjective facts). The latter is explained as a special case of the linguistic incommunicability of abilities to place oneself at will in a given state, such as the state of having lowered blood pressure, and the state of having wiggling cars.[12]

If you have a new experience, you gain abilities to remember and to imagine. After you taste Vegemite, and you learn what it's like, you can afterward remember the experience you had. By remembering how it once was, you can afterward imagine such an experience. Indeed, even if you eventually forget the occasion itself, you will very likely retain your ability to imagine such an experience.

Further, you gain an ability to recognize the same experience if it comes again. If you taste Vegemite on another day, you will probably know that you have met the taste once before. And if, while tasting Vegemite, you know that it is Vegemite you are tasting, then you will be able to put the name to the experience if you have it again. Or if you are told nothing at the time, but later you somehow know that it is Vegemite that you are then remembering or imagining tasting, again you can put the name to the experience, or to the memory, or to the experience of imagining, if it comes again. Here, the ability you gain is an ability to gain information if given other information. Nevertheless, the information gained is not phenomenal, and the ability to gain information is not the same thing as information itself.

Earlier, I mentioned 'knowing what an experience is like under a description.' Now I can say that what I meant by this was having the ability to remember or imagine an experience while also knowing the egocentric proposition that what one is then imagining is the experience of such-and-such description. One might well know what an experience is like under one

description, but not under another. One might even know what some experience is like, but not under any description whatever—unless it be some rather trivial description like 'that queer taste that I'm imagining right now.' That is what would happen if you slipped a dab of Vegemite into my food without telling me what it was: afterward, I would know what it was like to taste Vegemite, but not under that description, and not under any other non-trivial description. It might be suggested that 'knowing what it's like to taste Vegemite' really means what I'd call 'knowing what it's like to taste Vegemite under the description 'tasting Vegemite' ' ; and if so, knowing what it's like would involve both ability and information. I disagree. For surely it would make sense to say: 'I know this experience well, I've long known what it's like, but only today have I found out that it's the experience of tasting Vegemite.' But this verbal question is unimportant. For the information involved in knowing what it's like under a description, and allegedly involved in knowing what it's like, is anyhow not the queer phenomenal information that needs rejecting.

(Is there a problem here for the friend of phenomenal information? Suppose he says that knowing what it's like to taste Vegemite means knowing that the taste of Vegemite has a certain 'phenomenal character.' This requires putting the name to the taste, so clearly it corresponds to our notion of knowing what it's like to taste Vegemite under the description 'tasting Vegemite.' But we also have our notion of knowing what it's like *simpliciter,* and what can he offer that corresponds to that? Perhaps he should answer by appeal to a trivial description, as follows: knowing what it's like *simpliciter* means knowing what it's like under the trivial description 'taste I'm imagining now,' and that means knowing that the taste one is imagining now has a certain phenomenal character.)

As well as gaining the ability to remember and imagine the experience you had, you also gain the ability to imagine related experiences that you never had. After tasting Vegemite, you might for instance become able to imagine tasting Vegemite ice cream. By performing imaginative experiments, you can predict with some confidence what you would do in circumstances that have never arisen—whether you'd ask for a second helping of Vegemite ice cream, for example.

These abilities to remember and imagine and recognize are abilities you cannot gain (unless by super-neurosurgery, or by magic) except by tasting Vegemite and learning what it's like. You can't get them by taking lessons on the physics or the parapsychology of the experience, or even by taking comprehensive lessons that cover the whole of physics and parapsychology. The Ability Hypothesis says that knowing what an experience is like just *is* the possession of these abilities to remember, imagine, and recognize. It isn't the possession of any kind of information, ordinary or peculiar. It isn't knowing that certain possibilities aren't actualized. It isn't knowing-that. It's knowing-how. Therefore it should be no surprise that lessons won't teach you what an experience is like. Lessons impart information; ability is something else. Knowledge-that does not automatically provide know-how.

There are parallel cases. Some know how to wiggle their ears; others don't. If you can't do it, no amount of information will help. Some know how to eat with chopsticks, others don't. Information will help up to a point—for instance, if your trouble is that you hold one chopstick in each hand—but no amount of information, by itself, will bring you to a very high level of know-how. Some know how to recognize a C-38 locomotive by sight, others don't. If you don't, it won't much help if you memorize a detailed geometrical description of its shape, even though that does all the eliminating of possibilities that there is to be done. (Conversely, knowing the shape by sight doesn't enable you to write down the geometrical description.) Information very often contributes to know-how, but often it doesn't contribute enough. That's why music students have to practice.

Know-how is ability. But of course some aspects of ability are in no sense knowledge: strength, sufficient funds. Other aspects of ability are, purely and simply, a matter of information. If you want to know how to open the combination lock on the bank vault, information is all you need. It remains that there are aspects of ability that do *not* consist simply of possession of information, and that we *do* call knowledge. The Ability Hypothesis holds that knowing what an experience is like is that sort of knowledge.

If the Ability Hypothesis is the correct analysis of knowing what an experience is like, then phenomenal information is an illusion. We ought to explain that illusion. It would be feeble, I think, just to say that we're fooled by the ambiguity of the word 'know': we confuse ability with information because we confuse knowledge in the

sense of knowing-how with knowledge in the sense of knowing-that. There may be two senses of the word 'know,' but they are well and truly entangled. They mark the two pure endpoints of a range of mixed cases. The usual thing is that we gain information and ability together. If so, it should be no surprise if we apply to pure cases of gaining ability, or to pure cases of gaining information, the same word 'know' that we apply to all the mixed cases.

Along with information and ability, acquaintance is a third element of the mixture. If Lloyd George died too soon, there's a sense in which Father never can know him. Information won't do it, even if Father is a most thorough biographer and the archives are very complete. (And the trouble isn't that there's some very special information about someone that you can only get by being in his presence.) Know-how won't do it either, no matter how good Father may be at imagining Lloyd George, seemingly remembering him, and recognizing him. (Father may be able to recognize Lloyd George even if there's no longer any Lloyd George to recognize—if *per impossible* he did turn up, Father could tell it was him.) Again, what we have is not just a third separate sense of 'know.' Meeting someone, gaining a lot of information about him that would be hard to gain otherwise, and gaining abilities regarding him usually go together. The pure cases are exceptions.

A friend of phenomenal information will agree, of course, that when we learn what an experience is like, we gain abilities to remember, imagine, and recognize. But he will say that it is because we gain phenomenal information that we gain the abilities. He might even say the same about other cases of gaining know-how: you can recognize the C-38 when you have phenomenal information about what it's like to see that shape, you can eat with chopsticks or wiggle your ears when you gain phenomenal information about the experience of doing so, and so on. What should friends of the Ability Hypothesis make of this? Is he offering a conjecture, which we must reject, about the causal origin of abilities? I think not. He thinks, as we do, that experiences leave distinctive traces in people, and that these traces enable us to do things. Likewise being taught to recognize a C-38 or to eat with chopsticks, or whatever happens on first wiggling the ears, leave traces that enable us to do things afterward. That much is common ground. He also interprets these enabling traces as representations that bear information about their causes. (If the same traces had been caused in some deviant way they might perhaps have carried misinformation.) We might even be able to accept that too. The time for us to quarrel comes only when he says that these traces represent special phenomenal facts, facts which cannot be represented in any other way, and therefore which cannot be taught in physics lessons or even in parapsychology lessons. That is the part, and the *only* part, which we must reject. But that is no part of his psychological story about how we gain abilities. It is just a gratuitous metaphysical gloss on that story.

We say that learning what an experience is like means gaining certain abilities. If the causal basis for those abilities turns out also to be a special kind of representation of some sort of information, so be it. We need only deny that it represents a special kind of information about a special subject matter. Apart from that it's up for grabs what, if anything, it may represent. The details of stimuli: the chemical composition of Vegemite, reflectances of surfaces, the motions of well-handled chopsticks or of ears? The details of inner states produced by those stimuli: patterns of firings of nerves? We could agree to either, so long as we did not confuse 'having information' represented in this special way with having the same information in the form of knowledge or belief. Or we could disagree. Treating the ability-conferring trace as a representation is optional. What's essential is that when we learn what an experience is like by having it, we gain abilities to remember, imagine, and recognize.

NOTES

1. See Frank Jackson, "Epiphenomenal qualia," *Philosophical Quarterly* 32 (1982): pp. 127–36, and reprinted in this volume; "What Mary didn't know," *Journal of Philosophy* 83 (1986): pp. 291–95.
2. See B. A. Farrell, "Experience," *Mind* 59 (1950): pp. 170–98; and Thomas Nagel, "What is it like to be a bat?" *Philosophical Review* 83 (1974): pp. 435–50, also in Thomas Nagel, *Mortal Questions* (Cambridge, UK: Cambridge University Press, 1979).
3. See Peter Unger, "On experience and the development of the understanding," *American Philosophical Quarterly* 3 (1966): pp. 1–9.
4. For such speculation, see Paul M. Churchland, "Reduction, qualia, and the direct introspection of

brain states," *Journal of Philosophy* 82 (1985): pp. 8–28.

5. See my "Attitudes *de dicto* and *de se*," *Philosophical Review* 88 (1979): pp. 513–43, also in my *Philosophical Papers,* vol. I (New York: Oxford University Press, 1983); and Roderick Chisholm, *The First Person: An Essay on Reference and Intentionality* (Minneapolis: University of Minnesota Press, 1981).
6. See Gilbert Harman, "The intrinsic quality of experience," *Philosophical Perspectives* 4 (1990).
7. See Ned Block and Jerry A. Fodor, "What psychological states are not," *Philosophical Review* 81 (1972): pp. 159–81, and my "Mad pain and Martian pain," in *Readings in Philosophy of Psychology,* vol. I, Cambridge: Harvard University Press, 1983, and in my *Philosophical Papers,* vol. I, New York: Oxford University Press, 1983.
8. See my "New work for a theory of universals," *Australasian Journal of Philosophy* 61 (1983): pp. 343–77, especially pp. 361–64. For a different view about how to state minimal Materialism, see Terence Horgan, "Supervenience and microphysics," *Pacific Philosophical Quarterly* 63 (1982): pp. 29–43.
9. On recombination of possibilities, see my *On the Plurality of Worlds* (Oxford: Blackwell, 1986), pp. 87–92. The present argument may call for a principle that also allows recombination of properties; I now think that would not necessarily require treating properties as non-spatiotemporal parts of their instances. On recombination of properties, see also D. M. Armstrong, *A Combinatorial Theory of Possibility* (Cambridge: Cambridge University Press 1989).
10. See *On the Plurality of Worlds,* pp. 142–65, on linguistic ersatz possibilities.
11. On dependency hypotheses, see my "Causal decision theory," *Australasian Journal of Philosophy* 59 (1981): pp. 5–30, reprinted in *Philosophical Papers,* vol. II (New York: Oxford University Press, 1986).
12. Laurence Nemirow, "Review of Nagel's *Mortal Questions,*" *Philosophical Review* 89 (1980): pp. 475–76. For a fuller statement, see Nemirow, "Physicalism and the cognitive role of acquaintance," in *Mind and Cognition,* W. Lycan, ed., (Oxford: Blackwell, 1990); and Nemirow, "Functionalism and the Subjective Quality of Experience: doctoral dissertation, Stanford University, 1979. See also Michael Tye, "The subjective qualities of experience," *Mind* 95 (1986): pp. 1–17.

    I should record a disagreement with Nemirow on one very small point. We agree that the phrase 'what experience E is like' does not denote some 'subjective quality' of E, something which supposedly would be part of the subject matter of the phenomenal information gained by having E. But whereas I have taken the phrase to denote E itself, Nemirow takes it to be a syncategorematic part of the expression 'know what experience E is like.' See Nemirow, "Physicalism and the cognitive role of acquaintance" section III.

Naming and Necessity

Saul A. Kripke

. . . I finally turn to an all too cursory discussion of the application of the foregoing considerations to the identity thesis. Identity theorists have been concerned with several distinct types of identifications: of a person with his body, of a particular sensation (or event or state of having the sensation) with a particular brain state (Jones's pain at 06:00 was his C-fiber stimulation at that time), and of *types* of mental states with the corresponding *types* of physical states (pain is the stimulation of C-fibers). Each of these, and other types of identifications in the literature, present analytical problems, rightly raised by Cartesian critics, which cannot be avoided by a simple appeal to an alleged confusion of synonymy with identity. I should mention that there is of course no obvious bar, at least (I say cautiously) none which should occur to any intelligent thinker on a first reflection just before bedtime, to advocacy of some identity theses while doubting or denying others. For example, some philosophers have accepted the identity of particular sensations with particular brain states while denying the possibility of identities between mental and physical *types*.[1] I will concern myself primarily with the type-type identities, and the philosophers in question will thus be immune to much of the discussion; but I will mention the other kinds of identities briefly.

From *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980), pp. 144–55.

Descartes, and others following him, argued that a person or mind is distinct from his body, since the mind could exist without the body. He might equally well have argued the same conclusion from the premise that the body could have existed without the mind.[2] Now the one response which I regard as plainly inadmissible is the response which cheerfully accepts the Cartesian premise while denying the Cartesian conclusion, Let 'Descartes' be a name, or rigid designator, of a certain person, and let '*B*' be a rigid designator of his body. Then if Descartes were indeed identical to *B*, the supposed identity, being an identity between two rigid designators, would be necessary, and Descartes could not exist without *B* and *B* could not exist without Descartes. The case is not at all comparable to the alleged analogue, the identity of the first Postmaster General with the inventor of bifocals. True, this identity obtains despite the fact that there could have been a first Postmaster General even though bifocals had never been invented. The reason is that 'the inventor of bifocals' is not a rigid designator; a world in which no one invented bifocals is not *ipso facto* a world in which Franklin did not exist. The alleged analogy therefore collapses; a philosopher who wishes to refute the Cartesian conclusion must refute the Cartesian premise, and the latter task is not trivial.

Let '*A*' name a particular pain sensation, and let '*B*' name the corresponding brain state, or the brain state some identity theorist wishes to identify with *A*. Prima facie, it would seem that it is at least logically possible that *B* should have existed (Jones's brain could have been in exactly that state at the time in question) without Jones feeling any pain at all, and thus without the presence of *A*. Once again, the identity theorist cannot admit the possibility cheerfully and proceed from there; consistency, and the principle of the necessity of identities using rigid designators, disallows any such course. If *A* and *B* were identical, the identity would have to be necessary. The difficulty can hardly be evaded by arguing that although *B* could not exist without *A*, *being a pain* is merely a contingent property of *A*, and that therefore the presence of *B* without pain does not imply the presence of *B* without *A*. Can any case of essence be more obvious than the fact that *being a pain* is a necessary property of each pain? The identity theorist who wishes to adopt the strategy in question must even argue that *being a sensation* is a contingent property of *A*, for prima facie it would seem logically possible that *B* could exist without any sensation with which it might plausibly be identified. Consider a particular pain, or other sensation, that you once had. Do you find it at all plausible that *that very sensation* could have existed without being a sensation, the way a certain inventor (Franklin) could have existed without being an inventor?

I mention this strategy because it seems to me to be adopted by a large number of identity theorists. These theorists, believing as they do that the supposed identity of a brain state with the corresponding mental state is to be analyzed on the paradigm of the contingent identity of Benjamin Franklin with the inventor of bifocals, realize that just as his contingent activity made Benjamin Franklin into the inventor of bifocals, so some contingent property of the brain state must make it into a pain. Generally they wish this property to be one statable in physical or at least 'topic-neutral' language, so that the materialist cannot be accused of positing irreducible nonphysical properties. A typical view is that *being a pain*, as a property of a physical state, is to be analyzed in terms of the 'causal role' of the state,[3] in terms of the characteristic stimuli (e.g., pinpricks) which cause it and the characteristic behavior it causes. I will not go into the details of such analyses, even though I usually find them faulty on specific grounds in addition to the general modal considerations I argue here. All I need to observe here is that the 'causal role' of the physical state is regarded by the theorists in question as a contingent property of the state, and thus it is supposed to be a contingent property of the state that it is a mental state at all, let alone that it is something as specific as a pain. To repeat, this notion seems to me self-evidently absurd. It amounts to the view that the *very pain I now have* could have existed without being a mental state at all.

I have not discussed the converse problem, which is closer to the original Cartesian consideration—namely, that just as it seems that the brain state could have existed without any pain, so it seems that the pain could have existed without the corresponding brain state. Note that *being a brain state* is evidently an essential property of *B* (the brain state). Indeed, even more is true: not only being a brain state, but even being a brain state of a specific type is an essential property of *B*. The configuration of brain cells whose presence at a given time constitutes the presence of *B* at that time is essential to *B*, and in its absence *B* would not have existed. Thus someone who wishes to claim that the brain state and the pain are identical must

argue that the pain *A* could not have existed without a quite specific type of configuration of molecules. If *A* = *B*, then the identity of *A* with *B* is necessary, and any essential property of one must be an essential property of the other. Someone who wishes to maintain an identity thesis cannot simply *accept* the Cartesian intuitions that *A* can exist without *B*, that *B* can exist without *A*, that the correlative presence of anything with mental properties is merely contingent to *B*, and that the correlative presence of any specific physical properties is merely contingent to *A*. He must explain these intuitions away, showing how they are illusory. This task may not be impossible; we have seen above how some things which appear to be contingent turn out, on closer examination, to be necessary. The task, however, is obviously not child's play, and we shall see below how difficult it is.

The final kind of identity, the one which I said would get the closest attention, is the type-type sort of identity exemplified by the identification of pain with the stimulation of C-fibers. These identifications are supposed to be analogous with such scientific type-type identifications as the identity of heat with molecular motion, of water with hydrogen hydroxide, and the like. Let us consider, as an example, the analogy supposed to hold between the materialist identification and that of heat with molecular motion; both identifications identify two types of phenomena. The usual view holds that the identification of heat with molecular motion and of pain with the stimulation of C-fibers are both contingent. We have seen above that since 'heat' and 'molecular motion' are both rigid designators, the identification of the phenomena they name is necessary. What about 'pain' and 'C-fiber stimulation'? It should be clear from the previous discussion that 'pain' is a rigid designator of the type, or phenomenon, it designates: if something is a pain it is essentially so, and it seems absurd to suppose that pain could have been some phenomenon other than the one it is. The same holds for the term 'C-fiber stimulation,' provided that 'C-fibers' is a rigid designator, as I will suppose here. (The supposition is somewhat risky, since I know virtually nothing about C-fibers, except that the stimulation of them is said to be correlated with pain.[4] The point is unimportant; if 'C-fibers' is not a rigid designator, simply replace it by one which is, or suppose it used as a rigid designator in the present context.) Thus the identity of pain with the stimulation of C-fibers, if true, must be *necessary*.

So far the analogy between the identification of heat with molecular motion and pain with the stimulation of C-fibers has not failed; it has merely turned out to be the opposite of what is usually thought—both, if true, must be necessary. This means that the identity theorist is committed to the view that there could not be a C-fiber stimulation which was not a pain nor a pain which was not a C-fiber stimulation. These consequences are certainly surprising and counterintuitive, but let us not dismiss the identity theorist too quickly. Can he perhaps show that the apparent possibility of pain not having turned out to be C-fiber stimulation, or of there being an instance of one of the phenomena which is not an instance of the other, is an illusion of the same sort as the illusion that water might not have been hydrogen hydroxide, or that heat might not have been molecular motion? If so, he will have rebutted the Cartesian, not, as in the conventional analysis, by accepting his premise while exposing the fallacy of his argument, but rather by the reverse—while the Cartesian argument, given its premise of the contingency of the identification, is granted to yield its conclusion, the premise is to be exposed as superficially plausible but false.

Now I do not think it likely that the identity theorist will succeed in such an endeavor. I want to argue that, at least, the case cannot be interpreted as analogous to that of scientific identification of the usual sort, as exemplified by the identity of heat and molecular motion. What was the strategy used above to handle the apparent contingency of certain cases of the necessary *a posteriori*? The strategy was to argue that although the statement itself is necessary, someone could, *qualitatively* speaking, be in the same epistemic situation as the original, and in such a situation a *qualitatively* analogous statement could be false. In the case of identities between two rigid designators, the strategy can be approximated by a simpler one: Consider how the references of the designators are determined; if these coincide only contingently, it is this fact which gives the original statement its illusion of contingency. In the case of heat and molecular motion, the way these two paradigms work out is simple. When someone says, inaccurately, that heat might have turned out not to be molecular motion, what is true in what he says is that someone could have sensed a phenomenon in the same way we sense heat, that is, feels it by means of its production of the sensation we call 'the sensation of heat' (call it '*S*'), even though that phenomenon was

not molecular motion. He means, additionally, that the planet might have been inhabited by creatures who did not get *S* when they were in the presence of molecular motion, though perhaps getting it in the presence of something else. Such creatures would be, in some qualitative sense, in the same epistemic situation as we are, they could use a rigid designator for the phenomenon that causes sensation *S* in them (the rigid designator could even be 'heat'), yet it would not be molecular motion (and therefore not heat!), which was causing the sensation.

Now can something be said analogously to explain away the feeling that the identity of pain and the stimulation of C-fibers, if it is a scientific discovery, could have turned out otherwise? I do not see that such an analogy is possible. In the case of the apparent possibility that molecular motion might have existed in the absence of heat, what seemed really possible is that molecular motion should have existed without being *felt as heat,* that is, it might have existed without producing the sensation *S,* the sensation of heat. In the appropriate sentient beings is it analogously possible that a stimulation of C-fibers should have existed without being felt as pain? If this is possible, then the stimulation of C-fibers can itself exist without pain, since for it to exist without being *felt as pain* is for it to exist without there *being any* pain. Such a situation would be in flat out contradiction with the supposed necessary identity of pain and the corresponding physical state, and the analogue holds for any physical state which might be identified with a corresponding mental state. The trouble is that the identity theorist does not hold that the physical state merely *produces* the mental state, rather he wishes the two to be identical and thus *a fortiori* necessarily co-occurrent. In the case of molecular motion and heat there is something, namely, the sensation of heat, which is an intermediary between the external phenomenon and the observer. In the mental-physical case no such intermediary is possible, since here the physical phenomenon is supposed to be identical with the internal phenomenon itself. Someone can be in the same epistemic situation as he would be if there were heat, even in the absence of heat, simply by feeling the sensation of heat; and even in the presence of heat, he can have the same evidence as he would have in the absence of heat simply by lacking the sensation *S.* No such possibility exists in the case of pain and other mental phenomena. To be in the same epistemic situation that would obtain if one had a pain *is* to have a pain; to be in the same epistemic situation that would obtain in the absence of a pain *is* not to have a pain. The apparent contingency of the connection between the mental state and the corresponding brain state thus cannot be explained by some sort of qualitative analogue as in the case of heat.

We have just analyzed the situation in terms of the notion of a qualitatively identical epistemic situation. The trouble is that the notion of an epistemic situation qualitatively identical to one in which the observer had a sensation *S* simply *is* one in which the observer had that sensation. The same point can be made in terms of the notion of what picks out the reference of a rigid designator. In the case of the identity of heat with molecular motion the important consideration was that although 'heat' is a rigid designator, the reference of that designator was determined by an accidental property of the referent, namely the property of producing in us the sensation *S.* It is thus possible that a phenomenon should have been rigidly designated in the same way as a phenomenon of heat, with its reference also picked out by means of the sensation *S,* without that phenomenon being heat and therefore without its being molecular motion. Pain, on the other hand, is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality. Thus pain, unlike heat, is not only rigidly designated by 'pain' but the reference of the designator is determined by an essential property of the referent. Thus it is not possible to say that although pain is necessarily identical with a certain physical state, a certain phenomenon can be picked out in the same way we pick out pain without being correlated with that physical state. If any phenomenon is picked out in exactly the same way that we pick out pain, then that phenomenon *is* pain.

Perhaps the same point can be made more vivid without such specific reference to the technical apparatus in these lectures. Suppose we imagine God creating the world; what does He need to do to make the identity of heat and molecular motion obtain? Here it would seem that all He needs to do is to create the heat, that is, the molecular motion itself. If the air molecules on this earth are sufficiently agitated, if there is a burning fire, then the earth will be hot even if there are no observers to see it. God created light (and thus created streams of photons, according to present scientific doctrine) before He created human and animal observers; and

the same presumably holds for heat. How then does it appear to us that the identity of molecular motion with heat is a substantive scientific fact, that the mere creation of molecular motion still leaves God with the additional task of making molecular motion into heat? This feeling is indeed illusory, but what *is* a substantive task for the Deity is the task of making molecular motion felt as heat. To do this He must create some sentient beings to ensure that the molecular motion produces the sensation *S* in them. Only after he has done this will there be beings who can learn that the sentence 'Heat is the motion of molecules' expresses an a posteriori truth in precisely the same way that we do.

What about the case of the stimulation of C-fibers? To create this phenomenon, it would seem that God need only create beings with C-fibers capable of the appropriate type of physical stimulation; whether the beings are conscious or not is irrelevant here. It would seem, though, that to make the C-fiber stimulation correspond to pain, or be felt as pain, God must do something in addition to the mere creation of the C-fiber stimulation; He must let the creatures feel the C-fiber stimulation as *pain,* and not as a tickle, or as warmth, or as nothing, as apparently would also have been within His powers. If these things in fact are within His powers, the relation between the pain God creates and the stimulation of C-fibers cannot be identity. For if so, the stimulation could exist without the pain; and since 'pain' and 'C-fiber stimulation' are rigid, this fact implies that the relation between the two phenomena is not that of identity. God had to do some work, in addition to making the man himself, to make a certain man be the inventor of bifocals; the man could well exist without inventing any such thing. The same cannot be said for pain; if the phenomenon exists at all, no further work should be required to make it into pain.

In sum, the correspondence between a brain state and a mental state seems to have a certain obvious element of contingency. We have seen that identity is not a relation which can hold contingently between objects. Therefore, if the identity thesis were correct, the element of contingency would not lie in the relation between the mental and physical states. It cannot lie, as in the case of heat and molecular motion, in the relation between the phenomenon (= heat = molecular motion) and the way it is felt or appears (sensation *S*), since in the case of mental phenomena there is no 'appearance' beyond the mental phenomenon itself.

Here I have been emphasizing the possibility, or apparent possibility, of a physical state without the corresponding mental state. The reverse possibility, the mental state (pain) without the physical state (C-fiber stimulation) also presents problems for the identity theorists which cannot be resolved by appeal to the analogy of heat and molecular motion.

I have discussed similar problems more briefly for views equating the self with the body, and particular mental events with particular physical events, without discussing possible countermoves in the same detail as in the type-type case. Suffice it to say that I suspect that the considerations given indicate that the theorist who wishes to identify various particular mental and physical events will have to face problems fairly similar to those of the type-type theorist; he too will be unable to appeal to the standard alleged analogues.

That the usual moves and analogies are not available to solve the problems of the identity theorist is, of course, no proof that no moves are available. I certainly cannot discuss all the possibilities here. I suspect, however, that the present considerations tell heavily against the usual forms of materialism. Materialism, I think, must hold that a physical description of the world is a *complete* description of it, that any mental facts are 'ontologically dependent' on physical facts in the straight-forward sense of following from them by necessity. No identity theorist seems to me to have made a convincing argument against the intuitive view that this is not the case.[5]

NOTES

Part of this paper derives from a lecture at LaTrobe University in 1981. I thank LaTrobe for support in 1981, Harvard University for support under a Santayana Fellowship in 1988, and Frank Jackson for very helpful discussion.

1. Thomas Nagel and Donald Davidson are notable examples. Their views are very interesting, and I wish I could discuss them in further detail. It is doubtful that such philosophers wish to call themselves 'materialists.' Davidson, in particular, bases his case for his version of the identity theory on the supposed *impossibility* of correlating psychological properties with physical ones.

The argument against token-token identification in the text *does* apply to these views.

2. Of course, the body *does* exist without the mind and presumably without the person, when the body is a corpse. This consideration, if accepted, would already show that a person and his body are distinct. (See David Wiggins, "On Being at the Same Place at the Same Time," *Philosophical Review* 77 [1968]: pp. 90–95.) Similarly, it can be argued that a statue is not the hunk of matter of which it is composed. In the latter case, however, one might say instead that the former is 'nothing over and above' the latter; and the same device might be tried for the relation of the person and the body. The difficulties in the text would not then arise in the same form, but analogous difficulties would appear. A theory that a person is nothing over and above his body in the way that a statue is nothing over and above the matter of which it is composed, would have to hold that (necessarily) a person exists if and only if his body exists and has a certain additional physical organization. Such a thesis would be subject to modal difficulties similar to those besetting the ordinary identity thesis, and the same would apply to suggested analogues replacing the identification of mental states with physical states. A further discussion of this matter must be left for another place. Another view which I will not discuss, although I have little tendency to accept it and am not even certain that it has been set out with genuine clarity, is the so-called functional state view of psychological concepts.
3. For example, David Armstrong, *A Materialist Theory of the Mind*, London and New York, 1968, see the discussion review, "Armstrong on the Mind" by Thomas Nagel, *Philosophical Review* 79 (1970): pp. 394–403; and David Lewis, "An Argument for the Identity Theory," *The Journal of Philosophy*, 63: 1, 1966, pp. 17–25.
4. I have been surprised to find that at least one able listener took my use of such terms as 'correlated with,' 'corresponding to,' and the like as already begging the question against the identity thesis. The identity thesis, so he said, is not the thesis that pains and brain states are correlated, but rather that they are identical. Thus my entire discussion presupposes the anti-materialist position that I set out to prove. Although I was surprised to hear an objection which concedes so little intelligence to the argument, I have tried especially to avoid the term 'correlated' which seems to give rise to the objection. Nevertheless, to obviate misunderstanding, I shall explain my usage. Assuming, at least *arguendo,* that scientific discoveries have turned out so as not to refute materialism from the beginning, both the dualist and the identity theorist agree that there is a correlation or correspondence between mental states and physical states. The dualist holds that the 'correlation' relation in question is irreflexive; the identity theorist holds that it is simply a special case of the identity relation. Such terms as 'correlation' and 'correspondence' can be used neutrally without prejudging which side is correct.
5. Having expressed these doubts about the identity theory in the text, I should emphasize two things: first, identity theorists have presented positive arguments for their view, which I certainly have not answered here. Some of these arguments seem to me to be weak or based on ideological prejudices, but others strike me as highly compelling arguments which I am at present unable to answer convincingly. Second, rejection of the identity thesis does not imply acceptance of Cartesian dualism. In fact, my view above that a person could not have come from a different sperm and egg from the ones from which he actually originated implicitly suggests a rejection of the Cartesian picture. If we had a clear idea of the soul or the mind as an independent, subsistent, spiritual entity, why should it have to have any necessary connection with particular material objects such as a particular sperm or a particular egg? A convinced dualist may think that my views on sperms and eggs beg the question against Descartes. I would tend to argue the other way; the fact that it is hard to imagine me coming from a sperm and egg different from my actual origins seems to me to indicate that we have no such clear conception of a soul or self. In any event, Descartes' notion seems to have been rendered dubious ever since Hume's critique of the notion of a Cartesian self. I regard the mind-body problem as wide open and extremely confusing.

Acquaintance and the mind-body problem

Katalin Balog

> I wished to represent, in my own way, according to my own ideas, the material that was given to me, my material, myself . . . But there is something that I—perhaps understandably—didn't take into account: that we cannot ever represent ourselves *to* ourselves.[1]

In this chapter I will lay the groundwork for an account of *acquaintance* and discuss the consequences of the account for the metaphysics of mind. Acquaintance is a unique epistemological relation that relates a person to her own phenomenally conscious states and processes *directly*, *incorrigibly*, and *in a way that seems to reveal their essence*.[2] When one is aware of a phenomenal state in the process of having it, something essential about it is revealed, directly and incorrigibly—namely, *what it is like to have it*. Such an epistemic relation has struck many philosophers as deeply mysterious. One of the aims of this chapter is to dispel some of the mystery by providing an account of direct phenomenal concepts. These are the concepts deployed when a person is acquainted with her own conscious states in introspection, e.g., when I think to myself 'I have felt this in my shoulder before' upon noticing a familiar feeling as I throw a Frisbee. For reasons that will become clear I call my proposal 'the quotational account of direct phenomenal concepts.' The quotational account is a speculative proposal about human mental architecture. Although it is neutral between physicalist and dualist accounts of qualia in that both metaphysical views are compatible with it, if the general cognitive architecture invoked in it turns out to be correct physicalism scores a strategic victory. This is because the general cognitive architecture invoked in the quotational account has the resources to explain the nature of acquaintance. Therefore it obviates the need to explain acquaintance by way of appealing to the special, irreducibly mental, non-physical nature of phenomenal consciousness.

This chapter has two aims. The first is to elaborate on an account of phenomenal concepts that, in my view, yields a satisfying physicalist account of acquaintance. The second, related goal, is to show how such an account can be used as a powerful and quite general response to a whole slew of recent arguments against physicalism. I will start by briefly introducing these arguments.

1 Physicalism, Dualism, and the Zombie Argument

According to physicalism, the world's fundamental ontology is physical and the best account of that ontology is provided by fundamental physics. Contemporary physics tells us that this ontology consists of particles, strings, and fields of various types that occupy space-time (or bear spatiotemporal relations to one another) and possess a limited number of quantitative properties (mass, charge, electromagnetic potential, and so on). Physics also claims that there are only a few fundamental dynamical and perhaps non-dynamical laws that govern the structure of space-time and the evolution of its occupants. Physicalism thus understood is defined as follows: all truths, including truths about phenomenal consciousness, are metaphysically necessitated by the complete physical truth about the world.[3] This is the *Physicalist Entailment Thesis* (Phys):

(Phys) For all true statements T, $\Box (P \rightarrow T)$[4]

where *P* is the complete fundamental physical description of the world including the fundamental physical laws and also including a statement to the effect that it is complete.[5]

From *New Perspectives on Type Identity: The Mental and the Physical*, Simone Gozzano & Christopher S. Hill (eds.), pp. 16–43, Cambridge University Press, 2012.

If there are psychological truths—for example, that Mary knows what it is like to see red—that are not necessitated by *P* then physicalism is false.

According to dualism, the complete physical description of our world doesn't necessitate all mental truths: *P* leaves something out. Contemporary dualists generally do not think—as Descartes did—that what is left out are basic mental *entities* but they do maintain that there are basic mental (and proto-mental) *properties*, in particular, that there are basic phenomenal properties. They also usually think that there are fundamental laws that link phenomenal properties to certain properties of physical systems.[6] By their lights a complete description of our universe must include truths about where, when, and which conscious states are exemplified.

I won't rehearse the reasons to believe that physicalism is true.[7] But I will discuss some arguments that have persuaded many philosophers that physicalism is not true.[8]

There is a line of argument going back at least to Descartes' argument for the distinctness of mind and body that claims to show that physicalism is indeed false. In fact, these arguments can be understood to conclude, on the basis of a priori considerations, that *no* world where phenomenal properties are exemplified can be a purely physical world. The descendant of this argument that has received the most attention in the last decade is David Chalmers' 'Zombie Argument.'

Chalmers' most recent formulation of the zombie argument (Chalmers 2009) is as follows:

The zombie argument

(1) *P*&~*Q* is conceivable.[9]
(2) If *P*&~*Q* is conceivable then *P*&~*Q* is metaphysically possible (CP principle).
(3) If P&~Q is metaphysically possible then physicalism is false.

---

(4) Physicalism is false.

By 'statement *S* is conceivable' Chalmers 1996 means '*S* cannot be ruled out a priori.' Later Chalmers 2002 introduces a battery of conceivability concepts. For my present purposes I will bracket the complications that these different notions of conceivability introduce into the debate.[10]

The zombie argument is valid. Premise (3) is entailed by the proposition that Phys is a necessary condition for physicalism.[11] Philosophers who think that there is a functional or representational analysis of phenomenal consciousness reject (1).[12] But I agree with Chalmers that there is no functional or representational analysis of phenomenal consciousness and that no physical description a priori entails any positive phenomenal description. Later I will offer some considerations based on the nature of phenomenal concepts for why this is so, but for now I will just assume that (1) is true. So for both Chalmers and me the crucial premise in the argument is (2).[13]

How can physicalists respond to the zombie argument and its ilk? In Balog 1999 I refuted the zombie argument by arguing that if it is sound then it follows—given a few plausible assumptions—that a zombie counterpart to this argument is also sound. But it is not, hence the conceivability argument is unsound as well. However, in another paper (Balog, 'Illuminati, Zombies, and Metaphysical Gridlock,' unpublished MS) I show that Chalmers' zombie conceivability argument can be modified in a way that makes it resistant to this refutation. Here I propose to follow an approach—dubbed by Stoljar 2005 as the 'phenomenal concepts strategy'—that answers this new version of Chalmers' argument as well as other dualist arguments[14] by proposing a physicalist account of phenomenal concepts.

2 Desiderata for an Account of Phenomenal Concepts

Consciousness appears puzzling for many reasons—not just because of the conceivability of zombies. Below is a list of those features that seem most intractable for physicalism. I have gleaned these from the philosophical literature, but they also mostly strike me as what a non-philosopher would say, if not quite in these words. I suggest that a successful account of phenomenal concepts will explain these features, or most of them, since the traditional puzzles about consciousness are mostly *epistemic* in nature.

(1) *Only subjects who have undergone or are currently undergoing the relevant phenomenal states can token the*

corresponding phenomenal concepts. This underlies Jackson's 1982 knowledge argument and is widely accepted.[15]

(2) *Asymmetric epistemology. We are directly aware of our own conscious states in ways no one else can be.* One can be aware of one's conscious states simply by attending to them; to be aware of others' conscious states one has to observe their behavior. No one seems to contest this observation except Wittgensteinians and analytic behaviorists.

(3) *Transparency: when one turns one's attention to one's own conscious perceptual experience, one can become aware of the features of the objects perceived.* There is a stronger version of the transparency thesis advocated by representationalists.[16] Representationalists argue that when one attends to one's conscious experience, one is aware only of the representational content of the experience, or alternatively, only of features of the objects perceived, and conclude from this that qualia, i.e., intrinsic, qualitative, introspectible features of conscious experience, don't exist.[17]

(4) *Infallibility/incorrigibility: we seem to be infallible about certain judgments involving certain phenomenal concepts—e.g., my judging 'phenomenal red is occurring right now.'* The reason we tend to believe it is that it doesn't seem as though any belief concerning objective matters of fact can coherently override or correct our own judgment about what we feel when it occurs simultaneously with the experience.[18] I will argue that there are cases for which the thesis will come out true.[19]

(5) *Zombies are conceivable*, which means that the scenario in which zombies exist cannot be ruled out on a priori grounds. The main objectors to this are analytic functionalists.[20]

(6) *There is an explanatory gap.* No amount of knowledge about the physical facts (brain functioning and so on) is able to explain why a particular brain state/process has a particular feel, e.g. feels giddy. Whatever causal/functional/physical information we have about the brain processes that underlie phenomenal experience—i.e., about the neurophysiological, functional, or representational features of phenomenal experience—the fact that such experience has a distinct phenomenal character might still be *left out.* In contrast, all facts about water (that it is transparent, potable, etc.) are explicable in terms of facts about H_2O, together with physical and chemical laws. Nothing seems to be left out by such an explanation. Since we can't explain in the same way why a brain state feels giddy it is held that there is an explanatory gap between the phenomenal and the physical.

(7) *Acquaintance: we know our conscious states not by inference but by immediate acquaintance, which gives us direct, unmediated, substantial insight into their nature.* This, in opposition to the representationalist strong transparency thesis, commits one to the existence of qualia. I believe that qualia exist.[21] I think that we can attend to our experience and form direct, non-inferential concepts of its qualitative character that figure in phenomenal judgments. I also believe this gives us substantial insight into the nature of consciousness, and will shortly give an account of what this substantial insight consists in.

(8) *There is something it is like to have conscious states.* This, e.g., that there is something it is like to see a cloudless blue sky, is the most obvious 'given' about having conscious states. Even most representationalists don't question its existence. The denial of (8) qualifies one as an eliminativist about consciousness.

The task of the physicalist is to explain (1)–(8) in a manner compatible with physicalism. It is important to emphasize that this doesn't mean that the physicalist will have to give a perspicuous physical explanation of qualia; that is, close the explanatory gap. In my view once we understand what the explanatory gap consists in we will see that it cannot be closed. However, a satisfactory physicalist account should explain *this*, the fact that there is an unbridgeable explanatory gap, and show that all the other puzzling features of consciousness are, far from posing a problem for the physicalist view, features the physicalist will *expect* consciousness to have. Most theorists have attempted to explain (1)–(8) in terms of the nature of consciousness itself or to *explain away* these features. It is not surprising that neither physicalist nor dualist accounts of consciousness have been very successful at explaining these features since

features (1)–(7) are entirely *epistemic* features. So it seems reasonable to suppose that the key to their understanding will correspondingly lie in understanding the conceptual apparatus we use to think about them.

I propose focusing on our epistemic relation to consciousness, and especially on *acquaintance*, in trying to account for the puzzles of consciousness. This approach to the problems of consciousness has been aptly dubbed 'the phenomenal concept strategy.'[22]

3 The Constitutional Account of Phenomenal Concepts

I will assume in the following that concepts are or can be constituents of thoughts and that concepts and thoughts are representations. I will also assume that concepts are mental representations that are language-like—words of *Mentalese*.[23] The important point for the following is that since concepts and experiences are occurrent entities (events, states, processes) they can be constituents of one another and bear causal relations to one another.[24]

Concepts are the words of Mentalese. A particular token of a concept, e.g. DOG, possesses a number of different kinds of properties and relations that are relevant to my discussion: (i) realization properties, (ii) conceptual role, and (iii) semantic properties.

(i) When one tokens an instance of DOG, say in thinking the thought DOGS BARK, that token is realized by some neural state or process. The neural properties that are relevant to the token's being a token of DOG are its *realization properties*. A concept's realization properties are analogous to the particular physical type that realizes this (written or electronic) token of 'dog' or the particular sounds that realize a particular utterance of 'dog.'

(ii) A concept's *conceptual role* is the totality of causal relations (and dispositions) that tokens of thoughts containing the concept bear to each other and to perceptual inputs and behavioral outputs. Certain aspects of a concept's conceptual role may be essential to or even individuative of that concept while others are merely accidental; e.g., it is essential to the concept OR that one be inclined to make certain inferences, such as the inference from *P* to *PvQ*. It might also be essential to perceptual concepts, e.g. RED, that they be caused by certain perceptual inputs. Presumably, however, it is not essential to RED that one be caused to believe RED IS MY FAVORITE COLOR by the same perceptual inputs. How exactly to draw the distinction (which may be vague) between a concept's essential and non-essential roles is controversial.

(iii) A concept's *semantic features* concern what, if anything, the concept refers to. For example, the concept DOG refers to the property of being a dog. Exactly what determines the reference of a Mentalese word (with particular realization properties, syntax, and role) is a difficult and controversial matter. It is widely (though not universally) held that a concept's role (or the part of it essential to the concept) at least plays a part in determining the concept's reference. This part is the concept's *mode of presentation*. It often, but not always, has the form of a description—i.e., the thinker is disposed to infer the description from the tokening of the concept—e.g., from ARISTOTLE one is disposed to infer THE TEACHER OF ALEXANDER, etc. One can think of these descriptions as contents of a file attached to the concept. It is also widely accepted that reference is determined at least partly by external—causal, informational, or teleosemantic—relations of the concept to its environment.

A thinker typically has only partial epistemic access to features (i)–(iii) by introspection. When I attend to my thoughts I can typically obtain introspective knowledge of their semantic contents, e.g., that I am thinking about dogs. It is also plausible, though controversial, that one can obtain information about the conceptual roles of one's concepts—and which of these are essential—by intuitions based on thought experiments, e.g., by asking oneself questions like 'could one know p if p were false?' But the realization properties of one's Mentalese words—the 'shapes,' or 'mental ink' they are written in, so to speak—are almost always completely opaque. Almost always, with the exception—I propose—of *phenomenal concepts*.

I would like to propose an approach to phenomenal concepts that fits into this general framework and at the same time explains the

epistemic puzzles involving consciousness outlined above. An examination of the features of phenomenal concepts suggests that a successful account of phenomenal concepts will posit an intimate connection between conscious states and the concepts we form of them. Loar suggested the idea that phenomenal concepts are very special, direct demonstrative concepts.[25] Abstracting from some of the details, what he seems to have in mind is that when a person is having a particular experience she can deploy a concept that refers directly to the experience and that in some way the mode of presentation associated with the demonstrative involves the experience itself. How could we understand direct reference via these special modes of presentation? As Papineau points out, the suggestion doesn't help if by 'mode of presentation' we mean a description that we can already think and so we can use that description to think of an entity which has those properties.[26] That would be presupposing phenomenal concepts in the explanation of those very concepts. We have to think about the 'mode of presentation' of phenomenal concepts in some other way.

There is a problem with Loar's account that points the way towards an answer to our question above. Loar thinks of phenomenal concepts as in some way 'tracking' their referents. This suggests that he is thinking of the phenomenal concept and its referent as distinct entities related by causation. But it seems that this leaves too much of a distance between, e.g., a phenomenal concept P one applies to a particular pain **p**[27] as it occurs and **p** itself, as on this view their occurrence is independent. On a 'tracking' view, P, or rather, a concept just like P, could be tokened by someone in the complete absence of pain. A person like this would be a partial conceptual zombie; a *conceptual duplicate* of a normal human who, however, fails to have all the qualia the normal human has. But it seems to me that such a zombie is really impossible. Anybody who tokens a direct phenomenal concept as of a presently occurring pain is really in pain. The trouble with Loar's account is that it opens up the possibility of an appearance/reality distinction for direct phenomenal judgment whereas for direct phenomenal judgment there is no such distinction.

There is a way of thinking about phenomenal concepts which avoids these problems. It involves variations on the idea that (certain) phenomenal concepts are partly constituted by the phenomenal experiences they refer to.[28] On this view, a current phenomenal experience is *part* of the token concept currently applied to it, and the experience—at least partly—determines that the concept refers to the experience it contains. Of course, by 'part' I do not mean 'spatial part' but rather part in the sense that it is *metaphysically* impossible to token the concept without tokening its referent. I will cash this out presently. If this account is right, phenomenal concepts have very special *realization properties*: the neural states realizing these concepts are the very same neural states the concepts refer to!

This account is not intended to apply to all concepts that refer to phenomenal states or properties but only to 'direct phenomenal concepts.' But of course most of our reference to phenomenal states and qualia do not contain the phenomenal states themselves. What about 'indirect phenomenal concepts'? Clearly, a person can token a concept that refers to pain without her literally experiencing pain, as when she replies to her dentist's question by 'I am not in pain' or when one sees another person stub her toe and thinks THAT HURTS. Indirect phenomenal concepts are applied to non-occurrent (e.g., past or future) experiences of one's own or to the experiences of other people.[29] Understanding these is essential for understanding consciousness;[30] but for the rest of the chapter I will focus exclusively on direct phenomenal concepts.

Direct phenomenal concepts pick out their referent in virtue of their being partly constituted by a token of their reference.[31] In this they are unique among concepts. On this account, there is an intimate relation between a phenomenal concept and its referent; more intimate than any causal or tracking relation. It is also a way of cashing out the idea that the experience serves as its own mode of presentation.[32] The experience, so to speak, presents itself.

Later on I will fill in the details of my version of the constitutional account; but the core idea is what does the work in terms of explaining (1)–(7). Let me proceed to actually spell out those explanations.

(1) *Only subjects who have undergone or at least are currently undergoing the relevant phenomenal states can token the corresponding phenomenal concepts.* This is straightforwardly the case for direct phenomenal concepts because of the way they are constituted. In the case of indirect phenomenal concepts, the explanation is a bit more complicated.

(2) *Asymmetric epistemology*. One's awareness of one's own conscious states constitutively involves those very states. One couldn't be aware of another's states in the same way given the distinctness of the minds/ brains involved.

(3) *Transparency*: when one turns one's attention to one's own conscious perceptual experience, one is aware of the features of the objects perceived. On the constitutional account, the experience contained within the concept maintains its representational features; I take it that experiences including sensations, afterimages, phosphenes, etc. are representational.[33] So, for example, when a visual experience, i.e., a phenomenally conscious non-conceptual representation of an object (or objects) and their properties, partially constitutes a phenomenal concept representing it, attention directed to it will typically *also* or *primarily* be directed to the way the object is represented to be. I, however, deny the stronger version of the transparency thesis advocated by representationalists, namely the thesis that when one attends to one's conscious experience, one is aware *only* of the representational content of the experience. In my view, one can also direct one's attention to the phenomenal character of the experience, which is not identical to its representational content. (More on this in my explanation of acquaintance, under [7].)

(4) *Direct phenomenal judgments are infallible / incorrigible*. On the constitutional account, (4) will come out true for certain kinds of phenomenal judgments. For example, a phenomenal concept may refer to a particular type of visual experience, say the experience typically caused by seeing red objects in ordinary light, etc.—call this type of experience 'reddish'—by being constituted in part by a particular token of that type of experience. Then if I form the judgment I HAVE R where R is a direct phenomenal concept of reddish, my judgment cannot fail to be true.

This suggestion bears some similarity to Tyler Burge's 1988 account of self-knowledge. According to Burge, certain judgments about the intentional contents of one's states are self-certifying. Take for example, the judgment 'I am thinking that there may be life on another planet.' In order to make the judgment one has to do the thinking so the judgment must be true. From this point of view, Burge's account of our judgments about our thoughts, and the constitutional account of (certain of) our judgments about our experiences, are similar. Burge, however, doesn't offer any specific theory of our subjective concepts of our own thoughts. On my proposal, in order to token a direct phenomenal concept, one has to token the phenomenal state to which it refers, and *this is* what makes some of our phenomenal judgments self-certifying.[34]

(5) *The conceivability of zombies* is explained by the *directness* and *substantiality* of our direct phenomenal concepts, which, under the constitutional account, is compatible with physicalism. The directness of phenomenal concepts follows from the fact that the reference of a direct phenomenal concept is determined by how it is constituted and *not* by any description that is associated a priori with the concept. Phenomenal concepts are supposed to be different in this way, from concepts like WATER and even name concepts like CICERO. Chalmers and Jackson 2001 claim that these concepts are associated a priori with descriptions (e.g., 'the transparent potable liquid . . .'; 'the Roman orator who is at the origin of a causal chain culminating in this token'), and these connections are *sufficient* to rule out a priori a scenario where, e.g., everything is physically the same but yet there is no water. One doesn't have to commit to this to see that zombies are conceivable; however, the conceivability of zombies is only really significant if this is the case. So the point is that if one allows that this is true with respect to the concept WATER, or CICERO—thereby allowing the zombie argument to get off the ground—one *still* has to admit that it is not so with respect to phenomenal concepts; that the existence of zombies cannot be ruled out a priori. Because of the fundamentally different cognitive architecture of phenomenal concepts, there are no a priori connections between phenomenal and physical/ functional/ structural concepts that are sufficient to rule out a priori the zombie scenario.[35]

(6) *The explanatory gap*. Recall that the explanatory gap problem is that no amount of knowledge about the physical facts

(brain functioning and so on) is able to explain why a particular brain state/process has a particular feel, e.g., feels giddy. This contrasts with the way the fact that water is composed of H_2O molecules together with physical and chemical laws explains why water is potable, transparent, and so on. Once we have an explanation of why H_2O behaves in watery ways (and that it is the only substance that does so) we have an explanation of why water is H_2O. Since we can't explain why a brain state feels giddy in neurophysiological terms, we can't close the physical–phenomenal gap. You can see why this is in the following way. In the case of water and H_2O, the hypothesis that water = H_2O is quite natural in the light of all we know about H_2O and the laws that govern the behavior of H_2O—indeed, the opposite hypothesis doesn't even make sense. The hypothesis that the processes involving H_2O molecules are only nomologically correlated to the non-physical and non-chemical processes involving water is a non-starter.[36] On the other hand, the hypothesis that a phenomenal state is *identical* with a certain neurophysiological/functional state of the brain is just as compatible with our evidence as the opposing view. The hypothesis—endorsed by certain dualists—that phenomenal states and brain states are merely nomologically correlated makes perfect sense.

The difference is that while in the case of water we do not have any special access to its nature and properties that is not based on physical or functional information,[37] in the case of phenomenality we do. We do seem to have a special insight into the ultimate nature of phenomenal experience; and that nature doesn't seem captured or exhausted by any physical or functional description. As far as we know, that nature might elude any physical understanding. Notice that I stated the problem of the explanatory gap in a way that is independent of whether one subscribes to the semantic thesis discussed in the previous subsection that *all* but phenomenal terms have physical/functional analyses. It is significant that this can be done since it demonstrates that not all of the puzzles of consciousness will go away if we simply deny the semantic framework of the zombie argument. However, the constitutional account can explain why the explanatory gap arises, and it does so again in a way that is compatible with physicalism.

The constitutional account explains the gap by appealing to the direct and substantial grasp phenomenal concepts afford of their referent. When I focus on the phenomenal state, I have a 'substantive' grasp of its nature. I grasp it in terms of *what it's like* to be in that state. Because this grasp is substantive but at the same time independent of any causal or functional information (unlike in the case of WATER), information about the functioning of the brain simply won't explain *what it's like to be in that state*.

Since the issue of the substantive nature of phenomenal concepts is very closely connected to the issue of our acquaintance with phenomenal states, I'll take up the question of substantivity in the next section, together with the question of why the existence of the explanatory gap is not a problem for the physicalist.

(7) Acquaintance. We know our conscious states not by inference but by immediate acquaintance, which gives us direct, unmediated, substantial insight into their nature. If phenomenal concepts are partly constituted by phenomenal states, our knowledge of the presence of these states (in the first-person, subjective way of thinking of them) is not mediated by something distinct from these states. Rather the state itself serves as its own mode of presentation. Without getting deeply into philosophical issues involving perception, it is clear that this is quite different from visual (and other sensory) perception of external objects. On one account (with which I agree), when I visually perceive a red apple in front of me I token a phenomenal representation of the apple. The phenomenal representation and the apple are distinct existences and that at least leaves room for the possibility of illusion. When I focus on the phenomenal quality of that visual perception—not on what it represents but on the qualitative character of the visual experience—my representation contains that very experience. Thinking about it and simply having the experience will then share something very substantial,

very spectacular: namely the phenomenal character of the experience. And acquaintance, on this account, is the special, intimate epistemic relation we have to our phenomenal experience through the shared phenomenality of experience and thought. Shared phenomenality produces the sense that one has a direct insight into the nature of the experience. Hence the unique epistemic standing of acquaintance.

This last observation is connected with the explanatory gap. The core feature of phenomenal states that acquaintance reveals, i.e., their phenomenality, does not admit of explanation in terms of physical, functional, or structural features of brain states because of the very way we conceive of this feature, *directly, yet substantially* via acquaintance. Is this a problem for physicalism? You can see why not by focusing on what it means to have direct, unmediated insight into the nature of phenomenality. The important point is that this kind of direct insight (via shared phenomenality of thought and experience) does not reveal anything about the metaphysical nature of phenomenality. It is not the same sense of 'insight into the nature of X' as a scientific analysis of a brain state would provide. The one involves *having* the state; the other, analyzing it into its components. Those are very different activities. But there is a strong tendency to think that an insight into the nature of a phenomenon (e.g., via acquaintance) should lead one a priori to any other insights into the nature of the same phenomenon (e.g., via neuroscientific analysis), and so that any physical account of consciousness is thereby inadequate. This intuition also forms the basis of the conceivability arguments since Descartes and I believe it stems from a mistaken understanding of what it is to have a phenomenal insight into the nature of consciousness.

(8) *There is something it is like to have conscious states.* It should be clear by now that the constitutional account does not explain the phenomenality of brain states—it accepts and *explains* the existence of an explanatory gap between phenomenal and physical descriptions. The strategy is to show that all the epistemic features on our list, (1)–(7), can be accounted for by the special cognitive architecture involved in phenomenal concepts, and this special cognitive architecture is neutral with respect to the metaphysical nature of the phenomenal states involved. It is thus open to the physicalist to maintain that types of brain state are identical with types of phenomenal state. Of course there is no explanation of why this brain state type (neurophysiologically or functionally characterized) is identical with a phenomenal state type (phenomenally characterized)—hence the explanatory gap—but there is an explanation in terms of the constitutional account of why there is an explanatory gap even if physicalism is true. From this perspective, the puzzle that the explanatory gap presents is rather a trick the mind plays on itself as a result of the peculiar cognitive architecture involved in first-person phenomenal thought. This trick is, like a perceptual illusion, effective even in the face of intellectual conviction to the contrary. It is hard, even for the most devout physicalist, to shake the urge to get more of an explanation.[38] It is not unlike the urge, even after accepting Hume's demonstration that a non-question-begging justification of induction is not to be found, to still search for a justification.

Since the constitutional account is neutral about the nature of phenomenal properties, it can be adopted by a non-physicalist.[39] The explanations of most of the features will look much the same, with the exception that phenomenal concepts are constituted by non-physical states. However, there will be two explanations of why there is an unclosable explanatory gap. The dualist will say that the gap cannot be closed because phenomenal properties are not physical or functional properties. But this explanation is redundant since, as we have seen, the gap can also be explained merely in terms of direct phenomenal concepts.

4 The Quotational Account of Phenomenal Concepts

The constitutional account proposes that a certain kind of concept refers to something that (partly) constitutes it, and refers to it *in virtue of* it being so constituted but no actual account has been proposed of how a concept can be

like that. How can constitution determine reference? A dualist can attribute this to a primitive relation of acquaintance which doesn't itself require explanation. This seems to be an account of phenomenal reference by fiat. Can the physicalist do any better? Can we naturalize phenomenal self-reference?

The problem of naturalizing mental content is the problem of specifying the non-mental properties that determine the content of a particular concept, for example, specifying *in virtue of what* a particular concept refers to water. If the concept is complex, the question can be partially answered by an account of how the content of a concept with that structure is determined by the contents of its constituents. For simple concepts some other kind of account (or perhaps different accounts for different kinds of concepts) must be found. There have been a number of proposals;[40] all of them, in their present form, have problems.[41] I am not going to try to come up with a 'solution,' much less a general one. Rather, I will try to make it plausible that, in the particular case of direct phenomenal concepts, reference is determined by constitution. I will do this by showing that phenomenal concepts are analogous to quotation expressions and explaining how certain conceptual roles can make an operation mental quotation.

The question I want to shed light on then is this: why does a phenomenal concept (token) refer to a phenomenal experience that constitutes it, or, in the case of type phenomenal concepts, to the type of experience a token of which is constitutive of it, and most importantly, why does it so refer *in virtue* of this very fact of constitution? After all, this is not the case for most concepts. The concept DOG is not constituted by dogs, and the fact that the concept ATOM *is* constituted by atoms has nothing to do with why it refers to atoms. Information accounts and nomological accounts require an external relation between a concept and its referent unlike constitution, which makes them unsuitable candidates for the explanation of self-reference.[42] It seems plausible that one must look to the *conceptual role* of phenomenal concepts for an explanation of their self-referential nature.

The idea of an item partly constituting a representation that refers to that item is reminiscent of how linguistic quotation works. The referent of '—' is exemplified by whatever fills in the blank. In a quotation expression, a token of the referent is literally a constituent of the expression that refers to a type which it exemplifies and that expression has its reference (at least partly) in virtue of being so constituted. So, for example, 'dog' refers to the word spelled d-o-g, a token of which is enclosed between the quotation marks. Although in English we normally quote only expressions of English we can also quote foreign language representations and non-linguistic representations. We can even imagine, perhaps just as a joke, placing something which is not a representation, e.g., a cat, between quotes and thus producing a representation that everyone can understand refers to the type cat. My proposal is that there is a concept-forming mechanism that operates on an experience and turns it into a phenomenal concept that refers to either the token experience, or to a type of phenomenal experience that the token exemplifies. Further—and this is the heart of the proposal—the operation, like linguistic quotation, can be explained in terms of its conceptual roles.

A way to account for the semantics of quotation is to appeal to the disposition of competent language users to accept all instances of the disquotational schema on a priori grounds. So what accounts for the fact that ' ' is quotation in English is that users of '' who understand the meaning of 'refers,' etc., are disposed to accept all instances of the following schema on a priori considerations:

L1 'x' refers to x
L2 "x" refers to 'x'

where x stands in for any word of English. There is a potentially unlimited number of iterations of the schema at higher and higher levels.[43]

In presenting the *mental* disquotational schema, I can't simply offer a sentence schema in English, as I did with respect to linguistic quotation. To explicitly describe the Mentalese sentence schemas in question I will need to use special notation. In talking about Mentalese sentences, I will refer to concepts (Mentalese words) by CAPITALIZED WORDS as before, I will use '*' to refer to the mental quotation operation, and will use **bold font** to refer to the token experiences themselves that I claim to be part of these Mentalese sentences both inside and outside of the '*' operator. Notice in particular, that any expression in the position of '**experience x**,' for example, stands for an *experience*, and not a *concept* of Mentalese.

Here is my account of mental quotation. There is some mental operation (which I refer

to as '*') that takes an occurrent experience and forms it into a Mentalese concept referring to that experience.[44] What accounts for the semantics of this operation—i.e., that the resulting representation does refer to the very experience it is constituted by—is that competent thinkers who also have the concept of reference are disposed to accept all instances of the following schemas on a priori grounds:

M1 ***experience x*** REFERS-TO **experience x**
M2 ****experience x**** REFERS-TO ***experience x***

where '**experience x**' ranges over token experiences.[45]

Mental quotation, on this account, is analogous to linguistic quotation, with one difference. The difference is that, unlike linguistic quotation, what is between the mental quotes (*) at the first level is not a mental word but a mental representation that is not itself a word; it is an experience. I do think experiences represent—but not conceptually, so they are not in themselves concepts. This means that the expression on the right-hand side of M1 has simply experiences, i.e., non-conceptual representations as instances. Some might object that the resulting thoughts will not be well formed. However, I think there is a case to be made that plain—unquoted—experiences can be parts of thought under special circumstances.[46]

Let's take a closer look at an instance of M1. Suppose, for example, that you are currently having a visual experience of a patch of red, and that you are seeing it long enough to reflect on it. I propose that as you focus your attention on your experience you can form the thought

R1 ***reddish experience *e*** REFERS-TO **reddish experience *e***

where **reddish experience *e*** stands for a particular token reddish experience.

The rough 'translation' of this into English is the plausible claim that reddish experience refers to red. R1, however, is not in English; it is in Mentalese. And what my account requires is that all competent thinkers possessing phenomenal concepts and the concept REFERS have a disposition to accept R1, with its particular mode of presentation on a priori grounds. That it expresses a truth is not enough. The sentence "'dog' refers to canines,'" e.g., is true yet it is not the case that all possible competent speakers have a disposition to accept it on a priori grounds.

My claim is that R1 is just an expression of the *transparency thesis*, i.e., that when one turns one's attention to one's own conscious perceptual experience, it becomes evident that it represents the objects and/or features perceived.[47] I take the *transparency thesis* to be plausible on a priori grounds, and hence I take instances of M1, like, e.g., R1, to be compelling on a priori grounds.

How about the second-level mental disquotational schema, M2? Here the analogy with linguistic quotation is even closer. Both of the quotation expressions that appear in M2 are bona fide concepts. Considering our previous example again, I claim that one can reflect on one's direct phenomenal concept of a current reddish experience and realize that it refers to the very phenomenal character (reddish) that is phenomenally *present* in the concept. I suggest that the result of such reflection is the second-order judgment:

R2 ****reddish experience *e***** REFERS-TO ***reddish experience *e****.

This can be roughly expressed in English as the obvious thought that the concept of reddish refers to reddish. But, as before, notice that R2 is not in English, it is in Mentalese and the concept CONCEPT doesn't appear in it anywhere. Why believe that all possible competent thinkers—possessing phenomenal concepts and the concept REFERS—accept thoughts like this on a priori grounds *when they are conceived in the special way R2 affords*? I propose that this is simply explained by the awareness that phenomenal states are somehow 'present' in our concepts of them, embodied in the infallibility/incorrigibility intuition discussed above. The explanation of this awareness plausibly has to do with the nature of phenomenal states—but it is not my job to explore that here. All that matters for the present purposes is that such awareness exists, which I take as providing support for the view that all possible competent thinkers—possessing phenomenal concepts and the concept REFERS—have a disposition to accept instances of M2 on a priori grounds.[48]

You might have wondered by now about **reddish experience *e***: how can it appear twice in both R1 and R2? As I said, **reddish experience *e*** stands for a token experience; only token experiences can possibly be part of occurrent thoughts in Mentalese. But, just as the sentence "'red' refers to red'" involves two different tokens of the word 'red,' it seems that both R1 and R2 also involve two different tokens

of reddish experience. This is especially clear when you consider that mental quotation operates on a token experience and turns it into a token concept; a token experience 'taken up' into a token concept cannot be identical with a token experience that is not. In other words, **reddish experience *e***, ***reddish experience *e****, and ****reddish experience *e***** cannot all involve the same token experience but need to involve different tokens of the same type of experience. We can incorporate this in our schema in the following way:

M1+ ***experience x_1*** REFERS-TO **experience x_2**
M2+ ****experience x_1**** REFERS-TO ***experience x_2***

where '**experience x_1**' and '**experience x_2**' range over pairs of distinct experiences of the same type.

This seems to pose no problem for the quotational account. Just as in the linguistic case one of the necessary competencies of a speaker is to recognize tokens of the same word *as* tokens of the same word, one of the necessary competencies of a thinker in the mental case—involving phenomenal thought—is to recognize tokens of the same experience *as* tokens of the same experience if presented simultaneously or close to simultaneously. M1+ and M2+ will go through as long as **experience x_1** and **experience x_2** are both tokens of the same type and they are close enough in time—as they intuitively are if they appear in the same thought—for this to be evident for the subject entertaining the thought.

This completes my explanation of what makes the concept-forming mechanism that operates on phenomenal experience mental quotation. There is a further issue that I need to say more about. The reference of phenomenal type concepts *includes* the particular experience that constitutes the token of the concept, but will not be *exhausted* by it. A token of 'dog,' for example, includes in its reference the particular word between the quotes, but it might also refer to just all tokens of the word printed in lower-case type, or to all tokens written in any type or font, or to all tokens written or spoken, etc. Similarly, my phenomenal concepts 'reddish,' or 'dark-reddish,' or 'scarletish' can all be constituted by the same particular phenomenal experience; they might all be constituted as the concept ***experience *e**** where **experience *e*** happens to fall under all three concepts. What determines the type a phenomenal concept refers to, if the token experience that constitutes it doesn't, or at least doesn't fully, determine it? The quotational account is incomplete if it cannot answer this question.

I propose that the answer again has to do with conceptual roles. For example, what determines the reference of 'dog' on any particular occasion depends on the conceptual role of that instance of 'dog.' Both lower-case and capitalized versions of the word fall under this concept if, for example, were I to be confronted with some lower-case examples of the word and some examples of the word in capitals I would be inclined to judge 'same word.' The case is similar with phenomenal concepts. A particular token of the concept ***experience *e**** refers to, e.g., reddish experiences if, were I confronted with any kind of reddish experience, I would judge 'same kind of experience.' It refers to dark-reddish experiences if, were I confronted with dark-reddish experiences, I would judge 'same kind of experience' but not so when I am confronted with light-reddish experiences, etc.

5 Conclusion

The quotational account says that there is a cognitive mechanism that takes a phenomenal experience *e* and forms a phenomenal concept out of it that includes *e* in its reference. This cognitive mechanism is concept forming in this way *in virtue of* the conceptual roles of the resulting entities encoded in the schema M1 and M2. Further, what determines the scope of the concept is a further aspect of its conceptual role having to do with dispositions involving phenomenal similarity judgments. When we understand phenomenal concepts in this way the traditional puzzles of consciousness can be resolved.

It is important to see what this theory is not claiming. My theory is *not* that what it is to be a phenomenal state is to be mentally quoted. My view is the inverse, namely, that to be a phenomenal concept, a concept has to be constituted by a *phenomenal experience*. This means that *constitution matters* for phenomenal concepts. Phenomenal concepts are constituted by an instance of their referent, an experience with a phenomenal character, and we cannot help but be aware of the phenomenal character when we token the concept. This explains the sense that we are *acquainted* with phenomenal experience in a way that we are not acquainted with the referent of any other concept. It also explains the sense that these concepts seem to

allow us direct insight into the nature of their referent. Yet, nothing I have said in this chapter about phenomenal concepts is incompatible with physicalism; with the view that both phenomenal states and phenomenal concepts are realized by physical states. As a matter of fact, the supposition that phenomenal states are non-physical would add nothing to the explanatory power of the theory. I consider this a major argument for physicalism.

REFERENCES

Balog, K., "Conceivability, Possibility, and the Mind–Body Problem," *Philosophical Review* 108 (1999): pp. 497–528; reprinted in *The Philosopher's Annual*, vol. XXIII, P. Grim, K. Baynes, P. Ludlow, and G. Mar, eds., (Stanford, CA: Center for the Study of Language and Information.

_____ "The A Priori Entailment Thesis," *Philosophy and Phenomenological Research* 62 (2002): pp. 645–52. (An article on Frank Jackson's *From Metaphysics to Ethics*.)

_____ "Ontological Novelty, Emergence, and the Mind-Body Problem," in *Kreativität*, G. Abel, ed., (Hamburg: Meiner Verlag, 2006), pp. 26–45.

_____ "In Defense of the Phenomenal Concept Strategy," *Philosophy and Phenomenological Research* 84, no. 1 (2002).

Bealer, G., "Mental Properties," *Journal of Philosophy* 91 (1994): pp. 185–208.

Block, N., "Functional Role and Truth Conditions," *Proceedings of the Aristotelian Society* 61 (1987): pp. 157–81.

_____ "The Harder Problem of Consciousness," *Journal of Philosophy* 99 (2002): pp. 391–425.

_____ "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, M. Hahn and B. Ranberg, eds., (Cambridge, MA: MIT Press, 2003).

_____ "Max Black's Objection to Mind–Body Identity," in *Oxford Studies in Metaphysics*, II, D. Zimmerman, ed., (Oxford: Oxford University Press, 2006), pp. 3–78.

Block N., and Stalnaker, R., "Conceptual Analysis, Dualism, and the Explanatory Gap," *Philosophical Review* 108 (1999): pp. 1–46.

Burge, T., "Individualism and Self-Knowledge," *Journal of Philosophy* 85 (1988): pp. 649–63.

Chalmers, D. *The Conscious Mind* (New York: Oxford University Press, 1996).

_____ "Does Conceivability Entail Possibility?" in *Conceivability and Possibility*, T. Gendler and J. Hawthorne, eds., (New York: Oxford University Press, 2002), pp. 145–200.

_____ "The Content and Epistemology of Phenomenal Belief," in *Consciousness: New Philosophical Perspectives*, Q. Smith and A. Jokic, eds., (New York: Oxford University Press, 2003), pp. 220–73.

_____ "The Foundation of Two-Dimensional Semantics," in *Two-Dimensional Semantics: Foundations and Applications*, M. Garcia-Carpintero and J. Macià, eds., (New York: Oxford University Press, 2004), pp. 55–141.

_____ "Phenomenal Concepts and the Explanatory Gap," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, T. Alter and S. Walter, eds., (Oxford University Press, 2007), pp. 167–95.

_____ "The Two-Dimensional Argument against Materialism," in *The Oxford Handbook of the Philosophy of Mind*, B. McLaughlin, ed., (New York: Oxford University Press, 2009), pp. 313–39.

Chalmers, D., and Jackson, F., "Conceptual Analysis and Reductive Explanation," *Philosophical Review* 110 (2001): pp. 315–61.

Dretske, F. *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988).

Fodor, J. *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990).

Harman, G. "(Non-Solipsistic) Conceptual Role Semantics," in *New Directions in Semantics*, E. Lepore, ed., (London: Academic Press, 1987), pp. 55–81.

_____. "The Intrinsic Quality of Experience," in *Philosophical Perspectives 4 (1990), Action Theory and Philosophy of Mind*: pp. 31–52.

Jackson, F., "Epiphenomenal Qualia," *Philosophical Quarterly* 32 (1982): pp. 127–36.

_____ "Armchair Metaphysics," in *Philosophy in Mind*, M. Michael and John O'Leary-Hawthorne, eds., (Dordrecht: Kluwer, 1993), pp. 23–42.

_____ "Mind and Illusion," in *Minds and Persons*, A. O'Hear, ed., (Cambridge, UK: Cambridge University Press, 2003), pp. 251–71.

_____ "Representation and Experience," in *Representation in Mind: New Approaches to Mental Representation*, H. Clapin, P. Slezack, and P. Staines, eds., (New York: Elsevier, 2004), pp. 18–39.

Kirk, R. *Zombies and Consciousness* (New York: Oxford University Press, 2005).

Kripke, S. *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1972).

Levine, J. *Purple Haze* (New York: Oxford University Press, 2001).

_____ "Phenomenal Concepts and the Materialist Constraint," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, T. Alter and S. Walter, eds., (Oxford University Press, 2007), pp. 145–67.

Lewis, D. "An argument for the Identity Theory," *Journal of Philosophy* 63, no. 1 (1966): pp. 17–25.

_____ "New Work for a Theory of Universals," *Australasian Journal of Philosophy* 61 (1983): pp. 343–77.

Loar, B. 1990. "Phenomenal States," in *Philosophical Perspectives 4, Action Theory and Philosophy of Mind*: 81–108.

_____ "Phenomenal States," in *The Nature of Consciousness*, N. Block, O. Flanagan, and G. Güzeldere, eds., (Cambridge, MA: MIT Press, 1997), pp. 597–616 (revised version of Loar 1990).

_____ "Transparent Experience and the Availability of Qualia," in *Consciousness: New Philosophical Perspectives*, Q. Smith and A. Jokic, eds., (Oxford University Press, 2003), pp. 77–97.

Loewer, B., "An Argument for Strong Supervenience," *Supervenience: New Essays*, in E. Savellos and U. Yalcin, eds., (Cambridge, UK: Cambridge University Press, 1995), pp. 218–25.

_____ "From Physics to Physicalism," in *Physicalism and its Discontents*, K. Gillett and B. Loewer, eds., (Cambridge, UK: Cambridge University Press, 2001), pp. 37–57.

Martin, M. G. F., "The Transparency of Experience," *Mind and Language* 17 (2002): pp. 376–425.

McDowell, J., "The Content of Perceptual Experience," *Philosophical Quarterly* 44 (1994): pp. 190–206.

Melnyk, A. *A Physicalist Manifesto: Thoroughly Modern Materialism* (Cambridge, UK: Cambridge University Press, 2003).

Millikan, R., "Biosemantics," *Journal of Philosophy* 86 (1989): pp. 281–97.

Nagel, T., "What is it like to be a bat?" *Philosophical Review* 83 (1974): pp. 435–50.

Nida-Rümelin, M. "Grasping Phenomenal Properties," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, T. Alter and S. Walter, eds., (Oxford University Press, 2007), pp. 307–39.

Papineau, D., "Physicalism, Consciousness, and the Antipathetic Fallacy," *Australasian Journal of Philosophy* 71 (1993a): pp. 169–83.

_____ *Philosophical Naturalism* (Oxford: Blackwell Publishers, 1993b).

_____ "Arguments for Supervenience and Physical Realization," in *Supervenience: New Essays*, E. Savellos and U. Yalcin, eds., (Cambridge, UK: Cambridge University Press, 1995), pp. 226–42.

_____ *Thinking about Consciousness* (Oxford: Oxford University Press, 2002).

_____ "Phenomenal and Perceptual Concepts," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, T. Alter and S. Walter, eds., (Oxford University Press, 2007), pp. 111–45.

Robinson, H., "The Anti-Materialist Strategy and the Knowledge Argument," in *Objections to Physicalism*, Howard Robinson, ed., (Oxford: Oxford University Press, 1993), pp. 159–83.

Schwitzgebel, E., "The Unreliability of Naive Introspection," *Philosophical Review* 117 (2008): pp. 245–73.

Soames, S. *Reference and Description: The Case against Two-Dimensionalism* (Princeton, NJ: Princeton University Press, 2005).

Stoljar, D., "Two Conceptions of the Physical," *Philosophy and Phenomenological Research* 62 (2001): pp. 253–81.

_____ "The Argument from Diaphanousness," in *Language, Mind and World: Special Issue of the Canadian Journal of Philosophy*, M. Escurdia, R. Stanton, and C. Viger, eds., (Calgary, Alberta: University of Calgary Press, 2004), pp. 341–90.

_____ "Physicalism and Phenomenal Concepts," *Mind and Language* 20 (2005): pp. 469–95.

Sturgeon, S., "The Epistemic View of Subjectivity," *Journal of Philosophy* 91 (1994): pp. 221–36.

Tye, M. "Representationalism: The Theory and Its Motivations," in *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), pp. 45–68.

_____ *Consciousness Revisited: Materialism without Phenomenal Concepts* (Cambridge, MA: MIT Press, 2009).

White, S., "The Argument for the Semantic Premise," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, T. Alter and S. Walter, eds., (Oxford University Press, 2007), pp. 210–49.

Yablo, S., "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research* 53 (1993): pp. 1–42.

_____ "Coulda, Woulda, Shoulda," in *Conceivability and Possibility*, T. S. Gendler and J. Hawthorne, eds., (Oxford: Oxford University Press, 2002), pp. 441–92.

NOTES

Special thanks to David Papineau and Michael Della Rocca for very helpful comments on earlier drafts of this chapter. I would also like to thank Ned Block, Chris Hill, Joe Levine, Barry Loewer, Mike Martin, Gabriel Rabin, Howard Robinson, and audiences at the National Endowment for the Humanities (NEH) Summer Institute on Consciousness and Intentionality at University of California, Santa Cruz, 2002, the MIT Philosophy Department, the Summer Workshop of Collegium Budapest, 2004, and the Cognitive Science Group at the CUNY Graduate Center, 2010, for comments and criticism.

1. Imre Kertész, *A kudarc* (*Fiasco*) [1988] (Budapest: Magvetö Kiadó, 2003) 5th ed., p. 85. (Epigraph translated into English by Katalin Balog.)
2. I accept this with caveats—see the discussion later in this chapter.
3. This formulation is due to Jackson 1993. The first precise formulation of physicalism of this sort comes from Lewis 1983. Subsequent discussions are variations of the same theme. Many philosophers, among them non-physicalists, accept this kind of definition as capturing the intuitive notion of physicalism (see, e.g., Papineau 1993b; Chalmers 1996, pp. 41–42; Loewer 2001; Melnyk 2003).

4. More formally, the definition is: $(Y)(Y \rightarrow \Box (P \rightarrow Y))$, where Y is a sentential substitutional quantifier.
5. This last clause is needed to deal with the following complication in formulating physicalism. Statements that make reference to special kinds of property—to put it crudely, negative and global properties—are not necessitated by the full physical description of the world; they are only necessitated by the conjunction of the full physical description of the world together with the statement that it *is* the full fundamental description of the world. However, this issue will not make a difference for the rest of this chapter so I will ignore it.
6. I will assume that these laws are contingent; i.e., not metaphysically necessary. If laws are taken to be metaphysically necessary then it is difficult to state the difference between physicalism and dualism, since then both would hold that configurations of physical property instantiations metaphysically necessitate mental property instantiations.
7. For an argument for physicalism, see, e.g., Loewer 1995 and Papineau 1995.
8. In the empirical spirit recently gaining traction in philosophy, I would like to point out that according to a recent survey (conducted by David Bourget and David Chalmers in November 2009 at Philpapers, www.philpapers.org/surveys/), 27 percent of the sample—consisting mostly of professional philosophers, philosophy Ph.D. students, and some others—are dualists.
9. *P* is the complete fundamental physical description of the world, including the fundamental physical laws, and *Q* is a positive phenomenal truth, e.g., that someone is having a visual experience with a particular phenomenal character at a particular time.
10. Chalmers 2009 adds some clarifications and emendations to the argument. Since none of these affect my response to the zombie argument I will ignore them and stick with the simplified version of the argument.
11. Phys states that for all true positive statements *T*, $\Box (P \supset T)$, so if $P\&\sim Q$ is metaphysically possible then (Phys) is false and therefore physicalism is false.
12. For example Lewis 1966 and Jackson 2003.
13. Chalmers' defense and development of the two-dimensional framework and of the conceivability–possibility link can be found in Chalmers and Jackson 2001, Chalmers 2002, and Chalmers 2004. There are important discussions in Yablo 1993, 2002; Block and Stalnaker 1999; and Soames 2005. I briefly discuss what I think goes wrong with a related argument by Frank Jackson in Balog 2002.
14. Similar arguments include, among others, arguments based on conceivability considerations by Kripke 1972, Nagel 1974, Bealer 1994, and Chalmers 1996; 2009, as well as the knowledge argument of Jackson 1982, versions of the property dualism argument in Robinson 1993, White 2007, and Nida-Rümelin 2007, and the explanatory gap argument in Levine 2001 and 2007. The response to the zombie argument I offer via my account of phenomenal concepts can be adapted to respond to these other arguments as well, but in this chapter I will directly address only Chalmers' version.
15. For a denial of this claim, see Tye 2009.
16. See, e.g., Harman 1990, McDowell 1994, Tye 2000, and Jackson 2004 for transparency arguments.
17. For discussions of this argument see, e.g., Martin 2002 and Stoljar 2004.
18. Note the difference between this and perceptual illusions like the Müller-Lyer illusion. We cannot help but see the two lines as differing in length although we can correct the ensuing belief that they differ in length by, e.g., measuring them. On the other hand, no measurement, or, for that matter, no information about our brain states, would or should correct our judgment that the lines *appear* to be different in length.
19. Ryle and Wittgenstein were notable critics of the infallibility claim. See also Schwitzgebel 2008 for a rather pessimistic assessment of the reliability of introspective acquaintance with qualia.
20. See also Kirk 2005 for an interesting argument whose grounds go beyond analytic functionalism.
21. See Block 2003 and Loar 2003 for arguments for qualia. Dualists, naturally, tend to be committed to qualia.
22. The phenomenal concept strategy has been challenged by Stoljar 2005, Levine 2007, and Chalmers 2007. I respond to this challenge elsewhere (Balog 2012).
23. There may well be non-conceptual mental representations—image-like, map-like representations as well. It is plausible that tokens of phenomenal experience are non-conceptual representations.
24. There are philosophers who would like to avoid Mentalese or avoid representations altogether. It may be that my account can be made compatible with their ontologies but that is not something that I can do here.
25. See Loar 1990, 1997. The idea that the mind–body problem is a product of the special ways in which we conceive (in the first person) of our phenomenal states is first formulated in this chapter. A similar proposal by Scott Sturgeon 1994 appeals to the special epistemology of phenomenal states.
26. Papineau 2002, chapter 4.
27. The same problem, by and large, arises for type phenomenal concepts as well; however, because of complications having to do with failures of incorrigibility, I won't appeal to the type case here.
28. Similar ideas are proposed in Papineau 2002, Balog 2006, and Block 2006; Chalmers 2003 also puts forward a variation of this account.
29. The relationship between types of phenomenal concept and types of application is actually more complicated, as Kati Farkas has pointed it out to me. It is possible to apply direct phenomenal concepts to another's experience, as when one introspectively focuses on one's own experience of red and judges YOU ARE EXPERIENCING R, where R is a direct phenomenal concept formed on the basis of one's experience of red. However, the distinction between direct and indirect phenomenal concepts is not affected by this complication.
30. My view is that these concepts are individuated in part by conceptual roles that link them to direct phenomenal concepts.
31. There is a further complication. Direct phenomenal concepts, like the one I form of a buzzing sound as I listen to it, can refer either to particular (current) experiences of the thinker, or to phenomenal types exemplified in current conscious experience. I will indicate as I go which kind of concept I have in mind.
32. Some of Loar's remarks suggest that he might understand 'serves as its own mode of presentation' in this way, but other remarks suggest that he is thinking of the relation as causal.

33. I am not claiming that phenomenal experience can be analyzed in terms of or is exhausted by its representational character as representationalists hold but just that phenomenal experience purports to represent.
34. Notice that on Burge's view, judgments about our own experiences are not self-certifying in the way judgments about our own thoughts are. The judgment 'I have a reddish experience' is not self-certifying, at least not on the grounds that the judgment 'I am thinking that there might be life on other planets' is.
35. *Nota bene*: I am not denying that there are inferential links between thoughts involving direct phenomenal concepts that are individuative of them. I will argue that there are conceptual links between direct phenomenal concepts on the one hand, and indirect phenomenal concepts, other mental concepts, and behavioral concepts, etc., on the other. My point is that to the extent that these are a priori they are not of the sort that enables one to rule out a priori the zombie scenario.
36. Block and Stalnaker 1999 discuss the possibility of 'ghost water'—a non-physical kind that exists side by side with being composed of hydrogen and oxygen atoms and has all the same causal roles as the latter. Even if that is a coherent possibility, it would be the case that 'water' refers to both H_2O and ghost water and *not* that water refers to ghost water alone. So even in that possibility it wouldn't be the case that H_2O is merely nomologically connected to water.
37. Except for water's appearance properties, for example that its surface looks shiny in a storm, that it presents itself in a particular way to the touch, etc. But I am not going to press this point here.
38. Papineau 1993a, 2002, 2007 has an explanation he calls the 'anti-pathetic fallacy' which he uses to explain what he calls the 'intuition of distinctness,' that is, our intuition that physicalism cannot be right.
39. Chalmers 1996, 2003 suggests a constitutional account of phenomenal concepts.
40. For example informational accounts (Dretske 1988), nomological accounts (Fodor 1990), teleological accounts (Millikan 1989 and Papineau 1993b), and conceptual role accounts (Block 1987 and Harman 1987).
41. The inadequacy of physicalist accounts of content suggests that there may be an *explanatory gap* between the intentional and the physical as well as between the phenomenal and the physical. If there is such a gap, then it might be due to the failure of physicalism but it also might be due to the nature of the concepts we employ in attributing content.
42. Teleosemantics doesn't require external relations between a concept and its referent. Papineau 2002, 2007, who advocates a version of the constitutional account, appeals to teleosemantics to explain the reference of phenomenal concepts. But teleosemantics also holds (Papineau 2007) that the fact that phenomenal concepts are constituted by exemplars of their referent can play no direct role in explaining *why* they so refer. I disagree, as will be evident shortly.
43. This way of spelling out the concept-constituting conceptual role involves idealization. An ideal reasoner could use and understand any number of iterations of the quotation marks. In practice people won't be able to use or understand triple, quadruple, and higher-order quotation.
44. It is a superbly interesting—and at the moment wide-open question—exactly *what* this operation consists in. All that can be plausibly said on the basis of phenomenological evidence is that it involves attention of some sort.
45. In principle there might be further iterations of this schema at higher levels, as with linguistic quotation; but I doubt that our actual cognitive architecture allows us to form phenomenal concepts of phenomenal concepts anywhere beyond the third or fourth level.
46. R1 is controversial also on the grounds that it presupposes that all phenomenal experience represents. One could in fact omit R1 and establish the quotational account solely on the basis of R2; however, I believe considering R1 adds to the persuasiveness of the quotational account.
47. I would like to point out that this is not exactly what is typically advocated by representationalists when they appeal to the transparency of experience: R1 requires reflection *both* on phenomenal experience *and* the objects and/or properties it refers to. Representationalists think one can *only* reflect on the representational character of experience, i.e., on the external objects and their properties one's experience represents (and perhaps one's visual relation to these objects and properties), but *not* on the non-relational phenomenal features of experiences (see Loar 2003). As is clear by now, I disagree with this reading of transparency.
48. The sense in which instances of M1 and M2 are acceptable on a priori grounds is similar to the sense in which we can know a priori that we have phenomenal experience. It seems that the mere possession of phenomenal concepts is sufficient for knowledge of the existence of phenomenal states—though perhaps not in exactly the same way that possessing the concept BACHELOR is sufficient for knowing that bachelors are unmarried. I am not going to explore these issues further here.

Is Matter Conscious?

Hedda Hassel Mørch

The nature of consciousness seems to be unique among scientific puzzles. Not only do neuroscientists have no fundamental explanation for how it arises from physical states of the brain, we are not even sure whether we ever will. Astronomers wonder what dark matter is, geologists seek the origins of life, and biologists try to understand cancer—all difficult problems, of course, yet at least we have some idea of how to go about investigating them and rough conceptions of what their solutions could look like. Our first-person experience, on the other hand, lies beyond the traditional methods of science. Following the philosopher David Chalmers, we call it the hard problem of consciousness.

But perhaps consciousness is not uniquely troublesome. Going back to Gottfried Leibniz and Immanuel Kant, philosophers of science have struggled with a lesser known, but equally hard, problem of matter. What is physical matter in and of itself, behind the mathematical structure described by physics? This problem, too, seems to lie beyond the traditional methods of science, because all we can observe is what matter *does*, not what it *is* in itself—the 'software' of the universe but not its ultimate 'hardware.' On the surface, these problems seem entirely separate. But a closer look reveals that they might be deeply connected.

Consciousness is a multifaceted phenomenon, but subjective experience is its most puzzling aspect. Our brains do not merely seem to gather and process information. They do not merely undergo biochemical processes. Rather, they create a vivid series of feelings and experiences, such as seeing red, feeling hungry, or being baffled about philosophy. There is something that it's like to be you, and no one else can ever know that as directly as you do.

Our own consciousness involves a complex array of sensations, emotions, desires, and thoughts. But, in principle, conscious experiences may be very simple. An animal that feels an immediate pain or an instinctive urge or desire, even without reflecting on it, would also be conscious. Our own consciousness is also usually consciousness *of* something—it involves awareness or contemplation of things in the world, abstract ideas, or the self. But someone who is dreaming an incoherent dream or hallucinating wildly would still be conscious in the sense of having some kind of subjective experience, even though they are not conscious *of* anything in particular.

Where does consciousness—in this most general sense—come from? Modern science has given us good reason to believe that our consciousness is rooted in the physics and chemistry of the brain, as opposed to anything immaterial or transcendental. In order to get a conscious system, all we need is physical matter. Put it together in the right way, as in the brain, and consciousness will appear. But how and why can consciousness result merely from putting together non-conscious matter in certain complex ways?

This problem is distinctively hard because its solution cannot be determined by means of experiment and observation alone. Through increasingly sophisticated experiments and advanced neuroimaging technology, neuroscience is giving us better and better maps of what kinds of conscious experiences depend on what kinds of physical brain states. Neuroscience might also eventually be able to tell us what all of our conscious brain states have in common: for example, that they have high levels of integrated information (per Giulio Tononi's Integrated Information Theory), that they broadcast a message in the brain (per Bernard Baars' Global Workspace Theory), or that they generate 40-hertz oscillations (per an early proposal by Francis Crick and Christof Koch). But in all these theories, the hard problem remains. How and why does a system that integrates information, broadcasts a message, or oscillates at 40 hertz feel pain or delight? The appearance of consciousness from mere physical complexity seems equally mysterious no matter what precise form the complexity takes.

Nor would it seem to help to discover the concrete biochemical, and ultimately physical,

First published in *Nautilus*, 47: 90–97, 2017.

details that underlie this complexity. No matter how precisely we could specify the mechanisms underlying, for example, the perception and recognition of tomatoes, we could still ask: Why is this process accompanied by the subjective experience of red, or any experience at all? Why couldn't we have just the physical process, but no consciousness?

Other natural phenomena, from dark matter to life, as puzzling as they may be, don't seem nearly as intractable. In principle, we can see that understanding them is fundamentally a matter of gathering more physical detail: building better telescopes and other instruments, designing better experiments, or noticing new laws and patterns in the data we already have. If we were somehow granted knowledge of every physical detail and pattern in the universe, we would not expect these problems to persist. They would dissolve in the same way the problem of heritability dissolved upon the discovery of the physical details of DNA. But the hard problem of consciousness would seem to persist even given knowledge of every imaginable kind of physical detail.

In this way, the deep nature of consciousness appears to lie beyond scientific reach. We take it for granted, however, that physics can in principle tell us everything there is to know about the nature of physical matter. Physics tells us that matter is made of particles and fields, which have properties such as mass, charge, and spin. Physics may not yet have discovered all the fundamental properties of matter, but it is getting closer.

Yet there is reason to believe that there must be more to matter than what physics tells us. Broadly speaking, physics tells us what fundamental particles do or how they relate to other things, but nothing about how they *are* in themselves, independently of other things.

Charge, for example, is the property of repelling other particles with the same charge and attracting particles with the opposite charge. In other words, charge is a way of relating to other particles. Similarly, mass is the property of responding to applied forces and of gravitationally attracting other particles with mass, which might in turn be described as curving spacetime or interacting with the Higgs field. These are also things that particles do or ways of relating to other particles and to spacetime.

In general, it seems all fundamental physical properties can be described mathematically. Galileo, the father of modern science, famously professed that the great book of nature is written in the language of mathematics. Yet mathematics is a language with distinct limitations. It can only describe abstract structures and relations. For example, all we know about numbers is how they relate to the other numbers and other mathematical objects—that is, what they 'do,' the rules they follow when added, multiplied, and so on. Similarly, all we know about a geometrical object such as a node in a graph is its relations to other nodes. In the same way, a purely mathematical physics can tell us only about the relations between physical entities or the rules that govern their behavior.

One might wonder how physical particles *are*, independently of what they *do* or how they relate to other things. What are physical things like *in themselves*, or intrinsically? Some have argued that there is nothing more to particles than their relations, but intuition rebels at this claim. For there to be a relation, there must be two things being related. Otherwise, the relation is empty—a show that goes on without performers, or a castle constructed out of thin air. In other words, physical structure must be realized or implemented by some stuff or substance that is itself not purely structural. Otherwise, there would be no clear difference between physical and mere mathematical structure, or between the concrete universe and a mere abstraction. But what could this stuff that realizes or implements physical structure be, and what are the intrinsic, non-structural properties that characterize it? This problem is a close descendant of Kant's classic problem of knowledge of things-in-themselves. The philosopher Galen Strawson has called it the hard problem of matter.

It is ironic, because we usually think of physics as describing the hardware of the universe—the real, concrete stuff. But in fact physical matter (at least the aspect that physics tells us about) is more like software: a logical and mathematical structure. According to the hard problem of matter, this software needs some hardware to implement it. Physicists have brilliantly reverse-engineered the algorithms—or the source code—of the universe, but left out their concrete implementation.

The hard problem of matter is distinct from other problems of interpretation in physics. Current physics presents puzzles, such as: How can matter be both particle-like and wave-like? What is quantum wavefunction collapse? Are continuous fields or discrete individuals more fundamental? But these are all questions of how to properly conceive of the structure of reality.

The hard problem of matter would arise even if we had answers to all such questions about structure. No matter what structure we are talking about, from the most bizarre and unusual to the perfectly intuitive, there will be a question of how it is non-structurally implemented.

Indeed, the problem arises even for Newtonian physics, which describes the structure of reality in a way that makes perfect intuitive sense. Roughly speaking, Newtonian physics says that matter consists of solid particles that interact either by bumping into each other or by gravitationally attracting each other. But what is the intrinsic nature of the stuff that behaves in this simple and intuitive way? What is the hardware that implements the software of Newton's equations? One might think the answer is simple: It is implemented by solid particles. But solidity is just the behavior of resisting intrusion and spatial overlap by other particles—that is, another mere relation to other particles and space. The hard problem of matter arises for any structural description of reality no matter how clear and intuitive at the structural level.

Like the hard problem of consciousness, the hard problem of matter cannot be solved by experiment and observation or by gathering more physical detail. This will only reveal more structure, at least as long as physics remains a discipline dedicated to capturing reality in mathematical terms.

Might the hard problem of consciousness and the hard problem of matter be connected? There is already a tradition for connecting problems in physics with the problem of consciousness, namely in the area of quantum theories of consciousness. Such theories are sometimes disparaged as fallaciously inferring that because quantum physics and consciousness are both mysterious, together they will somehow be less so. The idea of a connection between the hard problem of consciousness and the hard problem of matter could be criticized on the same grounds. Yet a closer look reveals that these two problems are complementary in a much deeper and more determinate way. One of the first philosophers to notice the connection was Leibniz all the way back in the late 17th century, but the precise modern version of the idea is due to Bertrand Russell. Recently, contemporary philosophers including Chalmers and Strawson have rediscovered it. It goes like this.

The hard problem of matter calls for non-structural properties, and consciousness is the one phenomenon we know that might meet this need. Consciousness is full of qualitative properties, from the redness of red and the discomfort of hunger to the phenomenology of thought. Such experiences, or 'qualia,' may have internal structure, but there is more to them than structure. We know something about what conscious experiences are like *in and of themselves*, not just how they function and relate to other properties.

For example, think of someone who has never seen any red objects and has never been told that the color red exists. That person knows nothing about how redness relates to brain states, to physical objects such as tomatoes, or to wavelengths of light, nor how it relates to other colors (for example, that it's similar to orange but very different from green). One day, the person spontaneously hallucinates a big red patch. It seems this person will thereby learn what redness is like, even though he or she doesn't know any of its relations to other things. The knowledge he or she acquires will be non-relational knowledge of what redness is like in and of itself.

This suggests that consciousness—of some primitive and rudimentary form—is the hardware that the software described by physics runs on. The physical world can be conceived of as a structure of conscious experiences. Our own richly textured experiences implement the physical relations that make up our brains. Some simple, elementary forms of experiences implement the relations that make up fundamental particles. Take an electron, for example. What an electron does is to attract, repel, and otherwise relate to other entities in accordance with fundamental physical equations. What performs this behavior, we might think, is simply a stream of tiny electron experiences. Electrons and other particles can be thought of as mental beings with physical powers; as streams of experience in physical relations to other streams of experience.

This idea sounds strange, even mystical, but it comes out of a careful line of thought about the limitations of science. Leibniz and Russell were determined scientific rationalists—as evidenced by their own immortal contributions to physics, logic, and mathematics—but equally deeply committed to the reality and uniqueness of consciousness. They concluded that in order to give both phenomena their proper due, a radical change of thinking is required.

And a radical change it truly is. Philosophers and neuroscientists often assume that consciousness is like software, whereas the brain is

like hardware. This suggestion turns this completely around. When we look at what physics tells us about the brain, we actually just find software—purely a set of relations—all the way down. And consciousness is in fact more like hardware, because of its distinctly qualitative, non-structural properties. For this reason, conscious experiences are just the kind of things that physical structure could be the structure *of*.

Given this solution to the hard problem of matter, the hard problem of consciousness all but dissolves. There is no longer any question of how consciousness arises from non-conscious matter, because all matter is intrinsically conscious. There is no longer a question of how consciousness depends on matter, because it is matter that depends on consciousness—as relations depend on relata, structure depends on realizer, or software on hardware.

One might object that this is plain anthropomorphism, an illegitimate projection of human qualities on nature. After all, why do we think that physical structure needs some intrinsic realizer? Is it not because our own brains have intrinsic, conscious properties, and we like to think of nature in familiar terms? But this objection does not hold. The idea that intrinsic properties are needed to distinguish real and concrete from mere abstract structure is entirely independent of consciousness. Moreover, the charge of anthropomorphism can be met by a countercharge of human exceptionalism. If the brain is indeed entirely material, why should it be so different from the rest of matter when it comes to intrinsic properties?

This view, that consciousness constitutes the intrinsic aspect of physical reality, goes by many different names, but one of the most descriptive is 'dual-aspect monism.' Monism contrasts with dualism, the view that consciousness and matter are fundamentally different substances or kinds of stuff. Dualism is widely regarded as scientifically implausible, because science shows no evidence of any non-physical forces that influence the brain.

Monism holds that all of reality is made of the same kind of stuff. It comes in several varieties. The most common monistic view is physicalism (also known as materialism), the view that everything is made of physical stuff, which only has one aspect, the one revealed by physics. This is the predominant view among philosophers and scientists today. According to physicalism, a complete, purely physical description of reality leaves nothing out. But according to the hard problem of consciousness, any purely physical description of a conscious system such as the brain at least appears to leave something out: It could never fully capture what it is like to be that system. That is to say, it captures the objective but not the subjective aspects of consciousness: the brain function, but not our inner mental life.

Russell's dual-aspect monism tries to fill in this deficiency. It accepts that the brain is a material system that behaves in accordance with the laws of physics. But it adds another, intrinsic aspect to matter which is hidden from the extrinsic, third-person perspective of physics and which therefore cannot be captured by any purely physical description. But although this intrinsic aspect eludes our http://nautil.us/issue/47/consciousness/is-matter-conscious physical theories, it does not elude our inner observations. Our own consciousness constitutes the intrinsic aspect of the brain, and this is our clue to the intrinsic aspect of other physical things. To paraphrase Arthur Schopenhauer's succinct response to Kant: We can know the thing-in-itself because we are it.

Dual-aspect monism comes in moderate and radical forms. Moderate versions take the intrinsic aspect of matter to consist of so-called protoconscious or 'neutral' properties: properties that are unknown to science, but also different from consciousness. The nature of such neither-mental-nor-physical properties seems quite mysterious. Like the aforementioned quantum theories of consciousness, moderate dual-aspect monism can therefore be accused of merely adding one mystery to another and expecting them to cancel out.

The most radical version of dual-aspect monism takes the intrinsic aspect of reality to consist of consciousness itself. This is decidedly not the same as subjective idealism, the view that the physical world is merely a structure within human consciousness, and that the external world is in some sense an illusion. According to dual-aspect monism, the external world exists entirely independently of human consciousness. But it would not exist independently of any kind of consciousness, because all physical things are associated with some form of consciousness of their own, as their own intrinsic realizer, or hardware.

As a solution to the hard problem of consciousness, dual-aspect monism faces objections of its own. The most common objection is that it results in panpsychism, the view that all things are associated with some form of consciousness. To critics, it's just too implausible

that fundamental particles are conscious. And indeed this idea takes some getting used to. But consider the alternatives. Dualism looks implausible on scientific grounds. Physicalism takes the objective, scientifically accessible aspect of reality to be the only reality, which arguably implies that the subjective aspect of consciousness is an illusion. Maybe so—but shouldn't we be more confident that we are conscious, in the full subjective sense, than that particles are not?

A second important objection is the so-called combination problem. How and why does the complex, unified consciousness of our brains result from putting together particles with simple consciousness? This question looks suspiciously similar to the original hard problem. I and other defenders of panpsychism have argued that the combination problem is nevertheless not as hard as the original hard problem. In some ways, it is easier to see how to get one form of conscious matter (such as a conscious brain) from http://nautil.us/issue/47/consciousness/is-matter-conscious another form of conscious matter (such as a set of conscious particles) than how to get conscious matter from non-conscious matter. But many find this unconvincing. Perhaps it is just a matter of time, though. The original hard problem, in one form or another, has been pondered by philosophers for centuries. The combination problem has received much less attention, which gives more hope for a yet undiscovered solution.

The possibility that consciousness is the real concrete stuff of reality, the fundamental hardware that implements the software of our physical theories, is a radical idea. It completely inverts our ordinary picture of reality in a way that can be difficult to fully grasp. But it may solve two of the hardest problems in science and philosophy at once.

Content

INTRODUCTION

A ubiquitous feature of mental states is that they have *content*: that is, they represent features of the world. When I see a tree, my perceptual state represents the tree. When I believe that the Earth is round, my belief represents a state of the Earth. This feature of mental states is often called *intentionality*, or *aboutness*. A belief that Russell was a philosopher is about Russell, a desire to go to heaven is about heaven, and so on. A central feature of this sort of representation is that it can be assessed for *correctness*: my perception of water on the road ahead may be accurate or inaccurate; my belief that it is hot outside may be true or false; my desire for happiness may be satisfied or unsatisfied.

Mental content immediately raises a number of problems. How is it possible for one state of the world to represent another? How can a feature of the world to be correct or incorrect? Is it possible to account for mental content in physical terms, and if so how? What is the nature of the states by which we represent the world, especially beliefs and desires? How is the content of our mental states determined? Does this content depend on features internal to the subject, in the environment, or both?

A. The Nature of Intentionality

Frank Brentano, in his 1874 book *Psychology from an Empirical Standpoint*, famously held that intentionality is the mark of the mental. In the selection reproduced here (chapter 35), he raises the question of what distinguishes mental from physical phenomena. He canvases a number of possible answers - that mental phenomena are nonspatial, that they are objects of awareness, and so on—and settles on the claim that mental phenomena exhibit *intentional inexistence*: that is, they contain an *intentional object* within themselves, an object at which they are directed. Physical phenomena are never directed at an intentional object, according to Brentano, but mental phenomena are always directed at such an object.

Roderick Chisholm (chapter 36) takes his cue from Brentano, aiming to explore the idea of intentional inexistence in more detail. He focuses especially on the idea that the intentional object of a state may not actually exist, as when one believes that there are unicorns, and also on the idea that it is possible for two different states to be directed at the same object. This characterization goes beyond Brentano's, but Chisholm expresses a Brentano-like thesis by saying that descriptions of psychological phenomena involve object-directedness with these features, but descriptions of nonpsychological phenomena do not. He considers three ways in which intentionality might be accounted for in

simpler terms—by appeal to linguistic behavior, sign behavior, and expectation—and argues that each of these either presupposes psychological intentionality, or cannot account for the phenomena at issue. The conclusion seems to be that intentionality cannot be explained in non-psychological, non-intentional terms.

Fred Dretske's paper (chapter 37) can be seen as responding to the Brentano/Chisholm challenge, examining a way in which the intentionality of mental states might be grounded in the simple nonpsychological phenomena. Dretske argues that the central features of intentionality are present in a system as simple as a compass, which indicates the direction of the North Pole. He uses this sort of example to support a *causal theory* of content, on which a system represents features of its environment when it is causally connected to those features in the right sort of way. One question for such a theory is how a system can *misrepresent* a feature of its environment, but Dretske argues that an appeal to the natural function of the system (grounded in the system's history) can solve this problem. At the end of this paper, he discusses how this sort of intentionality can be extended into the sort of intentionality exhibited by thought.

Kathleen Akins (chapter 38) criticizes causal theories of representations such as Dretske's, by arguing that they do not work when applied to many sensory systems in the natural world. On the standard causal picture, sensory states are reliably caused by objective features of the environment and thereby represent them. Akins argues that sensory states are usually *narcissistic*: they do not reflect features of the environment (e.g. whether it is hot or cold), but instead they reflect relations to the organism itself (e.g. whether a given part of the body is too hot or too cold for the organism's safety and other purposes). Akins argues that understanding sensory systems this way requires reorienting our theories of sensory representation, so that the role of sensory states in representing the external world is less important than their role in guiding action.

Ruth Millikan (chapter 39) responds to the challenge of naturalizing intentionality in a different but related way, giving an account of representation that is grounded in the evolutionary functioning of cognitive systems. On this account, the content of a representation is determined by *normal conditions* for proper use of that representation by cognitive systems. Here the normal conditions are grounded in the history of the species as the conditions responsible (via natural selection) for the system's presence in later members of the species. On this account, representation is grounded in evolutionary *teleology*, in combination with the way a representation is used by cognitive systems.

The last two papers explore quite different accounts of the roots of intentionality. Robert Brandom (chapter 40) gives an *inferentialist* account on which propositional content is grounded in inferential role: to a first approximation, a contentful state is one that plays the right sort of role in inference. Brandom distinguishes his inferentialism from more familiar representational approaches to intentionality, and outlines a number of varieties of inferentialism before defending his own. He also addresses a number of challenges to inferentialism, including the question of which inferences play the central role in constituting meaning, and the question of whether inferentialism leads to an unacceptable holism, and whether it can account for the compositionality of mental phenomena.

Finally, Terry Horgan and John Tienson (chapter 41) suggest that intentional content is grounded in *phenomenology*: that is, in the character of conscious experience. In a way, this is the reverse of the representationalist claim that phenomenology is grounded in representation. Horgan and Tienson begin by examining the intimate connections between intentionality and phenomenology in both perception and belief. They go on to argue that much intentional content is determined by phenomenology, since any two beings with the same phenomenology will share a great deal of intentional content, irrespective of their other properties. This leads to a very different perspective on intentionality: they conclude that theories (such those above) on which intentionality is grounded in connections to the environment are fundamentally incorrect, and they suggest that a reductive account of intentionality will be at least as hard as a reductive account of phenomenology.

FURTHER READING

Chisholm's interpretation of Brentano is criticized by McAlister 1976. Dretske 1981, 1991; Millikan 1983, 1995; and Brandom 1994, 1999, elaborate their accounts at book-length. The huge literature on theories of content is well-represented in Stich and Warfield 1994, which includes criticisms of Dretske's and Millikan's accounts, and alternative accounts of content by Fodor (causal), Block (conceptual-role), and Cummins (interpretational). Causal accounts are explored at length by Fodor 1990. Searle 1991 also argues that intentionality requires phenomenology, and Siewert 1998 gives detailed argument for the claim that phenomenology determines intentional content in both perception and belief. Neander 2017 and Shea 2018 are recent book-length accounts of teleological approach to intentionality. Mendelovici 2018 develops a detailed phenomenal approach to intentionality, and Williams 2020 develops an interpretivist approach.

Brandom, R. *Making it Explicit* (Cambridge, MA: Harvard University Press, 1994).

_____ *Articulating Reasons* (Cambridge, MA: Harvard University Press, 2000).

Dretske, F. *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981).

_____*Explaining Behavior* (Cambridge, MA: MIT Press, 1991).

Fodor, J. A. *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990).

McAlister, L., "Chisholm and Brentano on intentionality," *Review of Metaphysics* 28 (1974): pp. 328–38. Reprinted in *The Philosophy of Brentano*, L. McAlister, ed., (London: Duckworth, 1976).

Mendelovici, A. *The Phenomenal Basis of Intentionality* (New York: Oxford University Press, 2018).

Millikan, R. *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1983).

_____*White Queen Psychology and Other Essays for Alive* (Cambridge, MA: MIT Press, 1995).

Neander, K. *A Mark of the Mental* (Cambridge, MA: MIT Press, 2017).

Searle, J. R. *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1991).

Shea, N. *Representation in Cognitive Science* (New York: Oxford University Press, 2018).

Siewert, C. *The Significance of Consciousness* (Princeton, NJ: Princeton University Press, 1998).

Stich, S. P., and Warfield, F., eds., *Mental Representation* (Oxford, UK: Blackwell, 1994).

Williams, R. *The Metaphysics of Representation* (Oxford, UK: Oxford University Press, 2020).

B. Propositional Attitudes

What is the nature of the mental states by which we represent the worlds? Arguably the most important mental states of this sort are *propositional attitudes*, such as beliefs and desires. These states involve different attitude to propositions such as the proposition that it is raining outside: one might believe this proposition, or desire that it be true, or hope that it be true, and so on. Proposition attitudes (especially beliefs and related states) are also known as *thoughts*. The papers in this section give various perspective on the nature of thoughts.

In the enclosed excerpts (chapter 42) from his important paper 'Empiricism and the Philosophy of Mind,' Wilfred Sellars addresses the nature of thoughts and the process of ascribing thoughts to one another. He tells a fable about our mythical 'Rylean ancestors,' an illusion to Gilbert Ryle's behaviorism (chapter 8). These ancestors only ascribe behavior to each other and never thoughts. One day a genius, Jones, develops a theory on which people's utterances are caused by certain mental states involving inner speech. Sellars uses this 'myth' to suggest that the idea of thinking may emerge

not from simple introspection but from a theory that explains the behavior of others. His picture involving inner speech also illustrates the idea that thought may involve language in a central role.

Jerry Fodor (chapter 43) pursues this idea further, arguing that there is a *language of thought* —a sort of internal mental language—and that thoughts are relations to sentences in this mental language. Fodor first sets out a number of conditions that a theory of propositional attitudes should meet, and he then discusses Carnap's idea that thoughts might be relations to sentence of English (or of the thinkers language in general). He argues that English itself will not work for this purpose, but that a distinct mental language might work. This suggestion is put forward as an empirical hypothesis; if it is correct, then there is a very close analogy between thought and language.

Daniel Dennett (chapter 44) offers a very different perspective on the nature of propositional attitudes, and in particular beliefs. Dennett argues that for a system to be a believer, and to have a belief, is for the system to be interpretable in a certain way by someone who adopts the 'intentional stance' the stance of predicting a system's behavior using beliefs and desires. On this view, to have a belief is closely tied to exhibiting certain patterns of behavior that allow for the right sort of predictability. Dennett discusses the question of whether beliefs 'really' exist on this view, and opts for an intermediate view: there are objective patterns in behavior, but they are sometimes susceptible to multiple interpretations.

Paul Churchland (chapter 45) gives reasons to hold that propositional attitudes might not exist at all. Like Sellars, Churchland holds that beliefs and desires are entities postulated by a theory used to explain behavior: the commonsense theory known as *folk psychology*. Churchland canvases some reasons to think that this theory may be radically *false*: there are many things it cannot explain, and it may be replaced entirely by a better scientific theory. When a theory is radically false, the entities it postulates are eliminated (as with the phlogiston theory of fire). So if folk psychology is radically false, beliefs and desires do not exist. The resulting view is *eliminativism*, or *eliminative materialism*.

Tamar Gendler (chapter 46) argues that our ordinary picture of propositional attitudes needs to be supplemented by a new sort of propositional attitude: *alief*. Alief is like belief in some ways, but it is associative, automatic, arational, affect-laden, and action-generating. For example, when one has the automatic sense that one will fall from a high place even though one does not, this is alief. One's aliefs can often be discordant from one's beliefs, for example in unconscious associations that depart from what one believes. Gendler suggests that in the moment-to-moment control of action, aliefs may play a more important role than beliefs.

FURTHER READING

Sellars' paper is reprinted in book form (with controversial notes by Brandom) in Sellars 1999, and also (with more extensive and less controversial notes) in deVries and Triplett 2000. Book-length developments of some of the ideas in this section are given by Fodor 1975, 1987; Dennett 1978, 1987; and Churchland 1990. These ideas are criticized in the respective collections Loewer and Rey 1990, Dahlberg 1993, and McCauley 1989. Field 1978 and Harman 1973 present alternative versions of the language of thought view. A different eliminativist view is outlined by Stich 1983. The question of whether folk psychology is a theory is addressed by the papers in Davies and Stone 1995 and Greenwood 1991.

Dahlberg, B., ed., *Dennett and his Critics* (Oxford: Blackwell, 1990).

Davies, M., and Stone, T., eds., *Folk Psychology: The Theory of Mind Debate* (Oxford: Blackwell, 1995).

Dennett, D. C. *Brainstorms* (Cambridge, MA: MIT Press, 1978).

_____ *The Intentional Stance* (Cambridge, MA: MIT Press, 1987).

deVries, W., and Triplett, T., *Knowledge, Mind, and the Given* (Indianapolis: Hackett, 2000).

Field, H., "Mental representation," *Erkenntnis* 13 (1978): pp. 9–18. Reprinted in Stich, S. P., and Warfield, F., eds., *Mental Representation.*

Fodor, J. A. *The Language of Thought* (Cambridge, MA: Harvard University Press, 1975).

_____ *Psychosemantics* (Cambridge, MA: MIT Press, 1987).

Greenwood, J. D., ed., *The Future of Folk Psychology: Intentionality and Cognitive Science* (Cambridge, UK: Cambridge University Press, 1991).

Harman, G. *Thought* (Princeton, NJ: Princeton University Press, 1973).

Loewer, B., and Rey, G. *Meaning in Mind: Fodor and his Critics* (Oxford: Blackwell, 1990).

McCauley, R. *The Churchlands and their Critics* (Oxford: Blackwell, 1996).

Sellars, W. *Empiricism and the Philosophy of Mind* (Cambridge, MA: Harvard University Press, 1999).

Stich, S. P. *From Folk Psychology to Cognitive Science* (Cambridge, MA: MIT Press, 1983).

C. Internalism and Externalism

Is the content of our thoughts determined by the internal properties of a subject, by the environment, or both? Of course almost everyone agrees that our thoughts refer to objects in the world, and that they are affected by the state of the world. Still, the traditional view has been a sort of *internalism*, holding that the *content* of our thoughts is determined by factors internal to the subject, so that any two subjects who are internal duplicates will have thoughts with the same content. More recently, a number of philosophers have argued for *externalism*, holding that the content of our thoughts is often determined by the state of the environment, so that two internally identical subjects may have different thoughts if they are in different environments. This issue has given rise to much fertile debate.

Hilary Putnam's "The Meaning of 'Meaning'" (chapter 47) provided much of the stimulus for this debate. In this paper, Putnam argues that the meaning of many of our words is not 'in the head' but depends on the environment. He introduces the important *Twin Earth* thought-experiment, involving two duplicate subjects living on planets that are superficially identical but that contain different liquids in the oceans and lakes: one (Earth) contains H_2O, and the other (Twin Earth) contains XYZ. Putnam argues that XYZ is not water but something else (twin water), and that when these subjects say 'water,' they mean quite different things. If this is right, then meaning depends directly on the environment. Putnam restricts his conclusion to language, but it is easily extended to the contents of thought: for example, one can argue that one subject *believes* that there is water in the ocean, and the other does not (he believes that there is twin water in the environment). If this is right, it seems to follow that the contents of our thoughts depend on external factors.

Tyler Burge (chapter 48) argues for externalism in a related but different way. He considers duplicate subjects who live in different *social* environments, in which words (such as 'arthritis') are used differently. He argues that as a result, the subjects have different beliefs: one has beliefs about arthritis, and the other has beliefs about a different disease. The result is a kind of *social externalism*, according to which what we think depends on the character of our social community.

Andy Clark and I (chapter 49) make a case for a very different sort of externalism, involving the active coupling of an organism with its environment. In cases where parts of the environment are fully integrated with a cognitive system (as with a notebook that serves as a memory), those parts of the environment count as part of the cognitive

system, and help to constitute the mental states of the subject in question. Clark and I call this view *active externalism*, as opposed to the 'passive externalism' of Putnam and Burge, and argue that it coheres well with a recent body of work in cognitive science.

Brie Gertler (chapter 50) responds by arguing that the active externalist picture threatens to overextend the mind. Subjects will have beliefs that they cannot introspect, and they will end up performing actions that they were not directly involved with. Gertler suggests that to avoid these consequences, we should embrace a picture of the mind where only occurrent conscious states are truly mental. Extended states, and internal dispositional and nonconscious states, are not really mental at all. The result is a strong form of internalism: the mind is entirely in the head, and entirely conscious too. A closely related form of internalism is developed by Terry Horgan and John Tienson in chapter 41, where they argue that all intentionality is grounded in consciousness, and consciousness is independent of the environment.

Dan Zahavi and Shaun Gallagher (chapter 51) argue that the mind is not just external but *embodied*: the body plays a crucial role in constituting the mind. They reject a Cartesian picture where a disembodied brain in a vat could have the same mental states as an embodied mind. They appeal to phenomenological discussions by Merleau-Ponty and others to sketch a picture where the body deeply shapes the mind. They distinguish many roles that the body plays in shaping our experience and our action, and draw in research on disability and on virtual reality to show how changes in our embodiment can deeply affect our minds. The resulting view of the embodied mind pins down one leg of a broad anti-Cartesian approach that is now often known as 4E cognition: a view on which the mind is extended, embedded, enactive, and embodied.

FURTHER READING

Many important papers on internalism and externalism are collected in Pessin and Goldberg 1996. Burge 1982 extends Putnam's thought experiment to the mental realm. Internalist responses to these are presented by Chalmers 2002, Fodor 1987, Loar 1988, Searle 1983, and Segal 2000, and are critiqued by Hawthorne and Yli-Vakkuri 2018. Externalist positions are elaborated by Stalnaker 1999 and Wilson 1995. Horgan and Tienson (chapter 41) argue for a sort of internalism based on phenomenological considerations; Tye (chapter 27) gives an externalist view of the contents of phenomenology. Active externalist views are elaborated by Clark 2008, Haugeland 1995, and Hurley 1998, and are criticized by Adams and Aizawa 2008 and Rupert 2010. Menary 2012 is a collection of articles on the topic. Gallagher's and Zahavi's views on embodiment and other topics are developed in Gallagher 2006 and Gallagher and Zahavi 2012. Newen et al 2018 is a collection of articles on 4E cognition.

Adams, F., and Aizawa, K. *The Bounds of Cognition* (Oxford: Wiley-Blackwell, 2008).

Burge, T., "Other bodies," in *Thought and Object*, A. Woodfield, ed. (London: Oxford University Press, 1982).

Chalmers, D. J., "The components of content," in *Philosophy of Mind: Classical and Contemporary Readings* (New York: Oxford University Press, 2002), 1st ed.

Clark, S. *Supersizing the Mind* (New York: Oxford University Press, 2008).

Fodor, J. *Psychosemantics* (Cambridge, MA: MIT Press, 1987).

Gallagher, S. *How the Body Shapes the Mind* (New York: Oxford University Press, 2006).

Gallagher, S., and Zahavi, D. *The Phenomenological Mind* (New York: Routledge, 2012).

Haugeland, J., "Mind embodied and embedded," in *Mind and Cognition*, Y. Houng and J. Ho, eds., (Taipei: Academia Sinica, 1995). Reprinted in Haugeland, J. *Having Thought* (Cambridge, MA: Harvard University Press, 1998).

Hawthorne, J., and Yli-Vakkuri, J. *Narrow Content* (Oxford: Oxford University Press, 2018).

Hurley, S. *Consciousness in Action* (Cambridge, MA: Harvard University Press, 1998).

Loar, B., "Social content and psychological content," in *Contents of Thought*, R. Grimm and D. Merrill, eds., (Tucson: University of Arizona Press, 1987). Reprinted in Pessin, A., and Goldberg, S. The Twin Earth Chronicles: Twenty Years of Reflection on Putnam's "The Meaning of Meaning."

Ludlow, P., and Martin, N. *Externalism and Self-Knowledge* (Stanford, CA: CSLI Press, 1998).

Menary, R. *The Extended Mind* (Cambridge, MA: MIT Press, 2012).

Newen, A., De Bruin, L., and Gallagher, S. *The Oxford Handbook of 4E Cognition* (New York: Oxford University Press, 2018).

Pessin, A., and Goldberg, S. The Twin Earth Chronicles: Twenty Years of Reflection on Putnam's "The Meaning of Meaning" (Armonk, NY: M. E. Sharpe Inc., 1996).

Rupert, R. *Cognitive Systems and the Extended Mind* (New York: Oxford University Press, 2010).

Searle, J. R. *Intentionality* (Cambridge, UK: Cambridge University Press, 1983).

Segal, G. *A Slim Book about Narrow Content* (Cambridge, MA: MIT Press, 2000).

Stalnaker, R. *Context and Content* (New York: Oxford University Press, 1999).

Wilson, R. A. *Cartesian Psychology and Physical Minds: Individualism and the Sciences of the Mind* (Cambridge, UK: Cambridge University Press, 1995).

The Distinction between Mental and Physical Phenomena

Franz Brentano

1. All the data of our consciousness are divided into two great classes—the class of physical and the class of mental phenomena. We spoke of this distinction earlier when we established the concept of psychology, and we returned to it again in our discussion of psychological method. But what we have said is still not sufficient. We must now establish more firmly and more exactly what was only mentioned in passing before.

This seems all the more necessary since neither agreement nor complete clarity has been achieved regarding the delimitation of the two classes. We have already seen how physical phenomena which appear in the imagination are sometimes taken for mental phenomena. There are many other such instances of confusion. And even important psychologists may be hard pressed to defend themselves against the charge of self-contradiction. For instance, we encounter statements like the following: sensation and imagination are distinguished by the fact that one occurs as the result of a physical phenomenon, while the other is evoked by a mental phenomenon according to the laws of association. But then the same psychologists admit that what appears in sensation does not correspond to its efficient cause. Thus it turns out that the so-called physical phenomenon does not actually appear to us, and, indeed, that we have no presentation of it whatsoever—certainly a curious misuse of the term 'phenomenon'! Given such a state of affairs, we cannot avoid going into the question in somewhat greater detail.

2. The explanation we are seeking is not a definition according to the traditional rules of logic. These rules have recently been the object of impartial criticism, and much could be added to what has already been said. Our aim is to clarify the meaning of the two terms '*physical phenomenon*' and '*mental phenomenon*,' removing all misunderstanding and confusion concerning them. And it does not matter to us what means we use, as long as they really serve to clarify these terms.

To this end, it is not sufficient merely to specify more general, more inclusive definitions. Just as deduction is opposed to induction when we speak of kinds of proof, in this case explanation by means of subsumption under a general term is opposed to explanation by means of particulars, through examples. And the latter kind of explanation is appropriate whenever the particular terms are clearer than the general ones. Thus it is probably a more effective procedure to explain the term 'color' by saying that it designates the class which contains red, blue, green and yellow, than to do the opposite and attempt to explain 'red' by saying it is a particular kind of color. Moreover, explanation through particular definitions will be of even greater use when we are dealing, as in our case, with terms which are not common in ordinary life, while those for the individual phenomena included under them are frequently used. So let us first of all try to clarify the concepts by means of examples.

Every idea or presentation which we acquire either through sense perception or imagination is an example of a mental phenomenon. By presentation I do not mean that which is presented, but rather the act of presentation. Thus, hearing a sound, seeing a colored object, feeling warmth or cold, as well as similar states of imagination are examples of what I mean by this term. I also mean by it the thinking of a general concept, provided such a thing actually does occur. Furthermore, every judgement, every recollection, every expectation, every inference, every conviction or opinion, every doubt, is a mental phenomenon. Also to be included under this term is every emotion: joy, sorrow, fear, hope, courage, despair, anger, love, hate, desire, act

Excerpted from D. Terrell, A. Rancurello, and L. McAlister, trans.; L. McAlister, ed., *Psychology from an Empirical Standpoint* (Routledge, 1995). Reprinted with permission of the publisher.

of will, intention, astonishment, admiration, contempt, etc.

Examples of physical phenomena, on the other hand, are a color, a figure, a landscape which I see, a chord which I hear, warmth, cold, odor which I sense; as well as similar images which appear in the imagination.

These examples may suffice to illustrate the differences between the two classes of phenomena.

3. Yet we still want to try to find a different and a more unified way of explaining mental phenomena. For this purpose we make use of a definition we used earlier when we said that the term 'mental phenomena' applies to presentations as well as to all the phenomena which are based upon presentations. It is hardly necessary to mention again that by 'presentation' we do not mean that which is presented, but rather the presenting of it. This act of presentation forms the foundation not merely of the act of judging, but also of desiring and of every other mental act. Nothing can be judged, desired, hoped or feared, unless one has a presentation of that thing. Thus the definition given includes all the examples of mental phenomena which we listed above, and in general all the phenomena belonging to this domain.

It is a sign of the immature state of psychology that we can scarcely utter a single sentence about mental phenomena which will not be disputed by many people. Nevertheless, most psychologists agree with what we have just said, namely, that presentations are the foundation for the other mental phenomena. Thus Herbart asserts quite rightly, 'Every time we have a feeling, there will be something or other presented in consciousness, even though it may be something very diversified, confused and varied, so that this particular presentation is included in this particular feeling. Likewise, whenever we desire something . . . we have before our minds that which we desire.'[1] . . .

Accordingly, we may consider the following definition of mental phenomena as indubitably correct: they are either presentations or they are based upon presentations in the sense described above. Such a definition offers a second, more simple explanation of this concept. This explanation, of course, is not completely unified because it separates mental phenomena into two groups.

4. People have tried to formulate a completely unified definition which distinguishes all mental phenomena from physical phenomena by means of negation. All physical phenomena, it is said, have extension and spatial location, whether they are phenomena of vision or of some other sense, or products of the imagination, which presents similar objects to us. The opposite, however, is true of mental phenomena; thinking, willing and the like appear without extension and without spatial location.

According to this view, it would be possible for us to characterize physical phenomena easily and exactly in contrast to mental phenomena by saying that they are those phenomena which appear extended and localized in space. Mental phenomena would then be definable with equal exactness as those phenomena which do not have extension or spatial location. Descartes and Spinoza could be cited in support of such a distinction. The chief advocate of this view, however, is Kant, who explains space as the form of the intuition of the external sense. . . .

But even on this point there is no unanimity among psychologists, and we hear it denied for contradictory reasons that extension and lack of extension are characteristics which distinguish physical and mental phenomena.

Many declare that this definition is false because not only mental phenomena, but also many physical phenomena appear to be without extension. A large number of not unimportant psychologists, for example, teach that the phenomena of some, or even of all of our senses originally appear apart from all extension and spatial location. In particular, this view is quite generally held with respect to sounds and olfactory phenomena. . . .

Others, as we said, will reject this definition for the opposite reason. It is not so much the assertion that all physical phenomena appear extended that provokes them, but rather the assertion that all mental phenomena lack extension. According to them, certain mental phenomena also appear to be extended. Aristotle seems to have been of this opinion when, in the first chapter of this treatise on sense and sense objects he considers it immediately evident, without any prior proof, that sense perception is the act of a bodily organ.[2] Modern psychologists and physiologists sometimes express themselves in the same way regarding certain affects. They speak of feelings of pleasure or pain which appear in the external organs, sometimes even after the amputation of the limb and yet, feeling, like perception, is a mental phenomenon. Some authors even maintain that sensory appetites appear localized. This view is shared by the poet when he speaks, not, to be sure, of thought, but of rapture and longing which suffuse the heart and all parts of the body.

Thus we see that the distinction under discussion is disputed from the point of view of both physical and mental phenomena. Perhaps both of these objections are equally unjustified.[3] At any rate, another definition common to all mental phenomena is still desirable. Whether certain mental and physical phenomena appear extended or not, the controversy proves that the criterion given for a clear separation is not adequate. Furthermore, this criterion gives us only a negative definition of mental phenomena.

5. What positive criterion shall we now be able to provide? Or is there perhaps no positive definition which holds true of all mental phenomena generally? Bain thinks that in fact there is none.[4] Nevertheless, psychologists in earlier times have already pointed out that there is a special affinity and analogy which exists among all mental phenomena, and which physical phenomena do not share.

Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental)[5] inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on.[6]

This intentional in-existence is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves.

But here, too, we come upon controversies and contradiction. Hamilton, in particular, denies this characteristic to a whole broad class of mental phenomena, namely, to all those which he characterizes as feelings, to pleasure and pain in all their most diverse shades and varieties. With respect to the phenomena of thought and desire he is in agreement with us. Obviously there is no act of thinking without an object that is thought, nor a desire without an object that is desired. 'In the phenomena of Feelings—the phenomena of Pleasure and Pain—on the contrary, consciousness does not place the mental modification or state before itself; it does not contemplate it apart—as separate from itself—but is, as it were, fused into one. The peculiarity of Feeling, therefore, is that there is nothing but what is subjectively subjective; there is no object different from the self—no objectification of any mode of self.'[7] In the first instance there would be something which, according to Hamilton's terminology, is 'objective,' in the second instance something which is 'objectively subjective,' as in self-awareness, the object of which Hamilton consequently calls the 'subject-object.' By denying both concerning feelings, Hamilton rejects unequivocably all intentional in-existence of these phenomena.

In reality, what Hamilton says is not entirely correct, since certain feelings undeniably refer to objects. Our language itself indicates this through the expressions it employs. We say that we are pleased with or about something, that we feel sorrow or grieve about something. Likewise, we say: that pleases me, that hurts me, that makes me feel sorry, etc. Joy and sorrow, like affirmation and negation, love and hate, desire and aversion, clearly follow upon a presentation and are related to that which is presented.

One is most inclined to agree with Hamilton in those cases in which, as we saw earlier, it is most easy to fall into the error that feeling is not based upon any presentation: the case of pain caused by a cut or a burn, for example. But the reason is simply the same temptation toward this, as we have seen, erroneous assumption. Even Hamilton recognizes with us the fact that presentations occur without exception and thus even here they form the basis of the feeling. Thus his denial that feelings have an object seems all the more striking.

One thing certainly has to be admitted; the object to which a feeling refers is not always an external object. Even in cases where I hear a harmonious sound, the pleasure which I feel is not actually pleasure in the sound but pleasure in the hearing. In fact you could say, not incorrectly, that in a certain sense it even refers to itself, and this introduces, more or less, what Hamilton was talking about, namely that the feeling and the object are 'fused into one.' But this is nothing that is not true in the same way of many phenomena of thought and knowledge, as we will see when we come to the investigation of inner consciousness. Still they retain a mental inexistence, a Subject-Object, to use Hamilton's mode of speech, and the same thing is true of these feelings. Hamilton is wrong when he says that with regard to feelings everything is 'subjectively subjective'—an expression which

is actually self-contradictory, for where you cannot speak of an object, you cannot speak of a subject either. Also, Hamilton spoke of a fusing into one of the feeling with the mental impression, but when carefully considered it can be seen that he is bearing witness against himself here. Every fusion is a unification of several things; and thus the pictorial expression which is intended to make us concretely aware of the distinctive character of feeling still points to a certain duality in the unity.

We may, therefore, consider the intentional inexistence of an object to be a general characteristic of mental phenomena which distinguishes this class of phenomena from the class of physical phenomena.

6. Another characteristic which all mental phenomena have in common is the fact that they are only perceived in inner consciousness, while in the case of physical phenomena only external perception is possible. This distinguishing characteristic is emphasized by Hamilton.[8]

It could be argued that such a definition is not very meaningful. In fact, it seems much more natural to define the act according to the object, and therefore to state that inner perception, in contrast to every other kind, is the perception of mental phenomena. However, besides the fact that it has a special object, inner perception possesses another distinguishing characteristic: its immediate, infallible self-evidence. Of all the types of knowledge of the objects of experience, inner perception alone possesses this characteristic. Consequently, when we say that mental phenomena are those which are apprehended by means of inner perception, we say that their perception is immediately evident.

Moreover, inner perception is not merely the only kind of perception which is immediately evident; it is really the only perception in the strict sense of the word.[9] As we have seen, the phenomena of the so-called external perception cannot be proved true and real even by means of indirect demonstration. For this reason, anyone who in good faith has taken them for what they seem to be is being misled by the manner in which the phenomena are connected. Therefore, strictly speaking, so-called external perception is not perception. Mental phenomena, therefore, may be described as the only phenomena of which perception in the strict sense of the word is possible.

This definition, too, is an adequate characterization of mental phenomena. That is not to say that all mental phenomena are internally perceivable by all men, and so all those which someone cannot perceive are to be included by him among physical phenomena. On the contrary, as we have already expressly noted above, it is obvious that no mental phenomenon is perceived by more than one individual. At the same time, however, we also saw that every type of mental phenomenon is present in every fully developed human mental life. For this reason, the reference to the phenomena which constitute the realm of inner perception serves our purpose satisfactorily.

7. We said that mental phenomena are those phenomena which alone can be perceived in the strict sense of the word. We could just as well say that they are those phenomena which alone possess real existence as well as intentional existence. Knowledge, joy and desire really exist. Color, sound and warmth have only a phenomenal and intentional existence.

There are philosophers who go so far as to say that it is self-evident that phenomena such as those which we call physical phenomena *could not* correspond to any reality. . . .

I must confess that I am unable to convince myself of the soundness of this argument. It is undoubtedly true that a color appears to us only when we have a presentation of it. We cannot conclude from this, however, that a color cannot exist without being presented. Only if the state of being presented were contained in the color as one of its elements, as a certain quality and intensity is contained in it, would a color which is not presented imply a contradiction, since a whole without one of its parts is indeed a contradiction. But this is obviously not the case. . . .

It is not correct, therefore, to say that the assumption that there exists a physical phenomenon outside the mind which is just as real as those which we find intentionally in us, implies a contradiction. It is only that, when we compare one with the other we discover conflicts which clearly show that no real existence corresponds to the intentional existence in this case. And even if this applies only to the realm of our own experience, we will nevertheless make no mistake if in general we deny to physical phenomena any existence other than intentional existence.

8. There is still another circumstance which people have said distinguishes between physical and mental phenomena. They say that mental phenomena always manifest themselves serially, while many physical phenomena manifest themselves simultaneously. But people do not always mean the same thing by this assertion,

and not all of the meanings which it has been given are in accord with the truth. . . .

Indeed, we could, with more reason, make the opposite assertion, namely, that very often many mental phenomena are present in consciousness simultaneously, while there can never be more than one physical phenomenon at a time.

What is the only sense, then, in which we might say that a mental phenomenon always appears by itself, while many physical phenomena can appear at the same time? We can say this insofar as the whole multiplicity of mental phenomena which appear to us in our inner perception always appear as a unity, while the same is not true of the physical phenomena which we grasp simultaneously through the so-called external perception. As happens frequently in other cases, so here, too, unity is confused by many psychologists with simplicity; as a result they have maintained that they perceive themselves in inner consciousness as something simple. Others, in contesting with good reason the simplicity of this phenomenon, at the same time denied its unity. The former could not maintain a consistent position because, as soon as they described their inner life, they found that they were mentioning a large variety of different elements; and the latter could not avoid involuntarily testifying to the unity of mental phenomena. They speak, as do others, of an 'I' and not of a 'we' and sometimes describe this as a 'bundle' of phenomena, and at the other times by other names which characterize a fusion into an inner unity. When we perceive color, sound, warmth, odor simultaneously nothing prevents us from assigning each one to a particular thing. On the other hand, we are forced to take the multiplicity of the various acts of sensing, such as seeing, hearing, experiencing warmth and smelling, and the simultaneous acts of willing and feeling and reflecting, as well as the inner perception which provides us with the knowledge of all those, as parts of one single phenomenon in which they are contained, as one single and unified thing. We shall discuss in detail later on what constitutes the basis for this necessity. At that time we shall also present several other points pertaining to the same subject. The topic under discussion, in fact, is nothing other than the so-called unity of consciousness, one of the most important, but still contested, facts of psychology.

9. Let us, in conclusion, summarize the results of the discussion about the difference between mental and physical phenomena. First of all, we illustrated the specific nature of the two classes by means of *examples.* We then defined mental phenomena as *presentations* or as phenomena which are based *upon presentation;* all the other phenomena being physical phenomena. Next we spoke of *extension*, which psychologists have asserted to be the specific characteristic of all physical phenomena, while all mental phenomena are supposed to be unextended. This assertion, however, ran into contradictions which can only be clarified by later investigations. All that can be determined now is that all mental phenomena really appear to be unextended. Further we found that the *intentional inexistence*, the reference to something as an object, is a distinguishing characteristic of all mental phenomena. No physical phenomenon exhibits anything similar. We went on to define mental phenomena as the exclusive *object of inner perception;* they alone, therefore, are perceived with immediate evidence. Indeed, in the strict sense of the word, they alone are perceived. On this basis we proceeded to define them as the only phenomena which possess *actual existence* in addition to intentional existence. Finally, we emphasized as a distinguishing characteristic the fact that the mental phenomena which we perceive, in spite of all their multiplicity, *always* appear to us *as a unity*, while physical phenomena, which we perceive at the same time, do not all appear in the same way as parts of one single phenomenon.

That feature which best characterizes mental phenomena is undoubtedly their intentional inexistence. By means of this and the other characteristics listed above, we may now consider mental phenomena to have been clearly differentiated from physical phenomena. . . .

NOTES

1. *Psychologie als Wissenschaft*, Part II, Sect. 1, Chap. 1, No. 103. Cp. also Drobisch, *Empirische Psychologie*, p. 38, and others of Herbart's school.
2. *De Sensu et Sensibili*, 1, 436, b. 7. Cp. also what he says in *De Anima*, I, 1, 403, 16, about affective states, in particular about fear.
3. The assertion that even mental phenomena appear to be extended rests obviously on a confusion of mental and physical phenomena similar to the confusion which we became convinced of above when we pointed out that a presentation is also the necessary foundation of sensory feelings.

4. *The Senses and the Intellect*, Introduction.
5. They also use the expression 'to exist as an object (objectively) in something,' which, if we wanted to use it at the present time, would be considered, on the contrary, as a designation of a real existence outside the mind. At least this is what is suggested by the expression 'to exist immanently as an object,' which is occasionally used in a similar sense, and in which the term 'immanent' should obviously rule out the misunderstanding which is to be feared.
6. Aristotle himself spoke of this mental in-existence. In his books on the soul he says that the sensed object, as such, is in the sensing subject; that the sense contains the sensed object without its matter; that the object which is thought is in the thinking intellect. In Philo, likewise, we find the doctrine of mental existence and in-existence. However, since he confuses them with existence in the proper sense of the word, he reaches his contradictory doctrine of the *logos* and Ideas. The same is true of the Neoplatonists. St. Augustine in his doctrine of the *Verbum mentis* and of its inner origin touches upon the same fact. St. Anselm does the same in his famous ontological argument; many people have observed that his consideration of mental existence as a true existence is at the basis of his paralogism (cp. Überweg, *Geschichte der Philosophie*, II). St. Thomas Aquinas teaches that the object which is thought is intentionally in the thinking subject, the object which is loved in the person who loves, the object which is desired in the person desiring, and he uses this for theological purposes.
7. *Lecture on Metaphysics*, I, 432.
8. *Lecture on Metaphysics*, I, 432.
9. [Translators' note: The German word which we translate as 'perception' is '*Wahrnehmung*' which literally means taking something to be true. The English word does not reflect this literal meaning so this paragraph only makes sense if we bear in mind the German word.]

'Intentional Inexistence'

Roderick M. Chisholm

1

. . . Psychological phenomena, according to Brentano, are characterized 'by what the scholastics of the Middle Ages referred to as the intentional (also the mental) inexistence of the object, and what we, although with not quite unambiguous expressions, would call relation to a content, direction upon an object (which is not here to be understood as a reality), or immanent objectivity.'[1] This 'intentional inexistence,' Brentano added, is peculiar to what is psychical; things which are merely physical show nothing like it.

Assuming, or *accepting*, is one of the phenomena Brentano would have called intentional. I will first try to formulate Brentano's thesis somewhat more exactly; then I will ask whether it is true of assuming.

2

The phenomena most clearly illustrating the concept of 'intentional inexistence' are what are sometimes called psychological attitudes; for example, desiring, hoping, wishing, seeking, believing, and assuming. When Brentano said that these attitudes 'intentionally contain an object in themselves,' he was referring to the fact that they can be truly said to 'have objects' even though the objects which they can be said to have do not in fact exist. Diogenes could have looked for an honest man even if there hadn't been any honest men. The horse can desire to be fed even though he won't be fed. James could believe there are tigers in India, and *take* something there to be a tiger, even if there aren't any tigers in India.

But *physical*—or nonpsychological—phenomena, according to Brentano's thesis, cannot thus 'intentionally contain objects in themselves.' In order for Diogenes to sit in his tub, for example, there must be a tub for him to sit in; in order for the horse to eat his oats, there must be oats for him to eat; and in order for James to shoot a tiger, there must be a tiger there to shoot.

The statements used in these examples seem to have the form of relational statements. 'Diogenes sits in his tub' is concerned with a relation between Diogenes and his tub.

From *Perceiving: A Philosophical Study* (Cornell University Press, 1957).

Syntactically, at least, 'Diogenes looks for an honest man' is similar: Diogenes' quest seems to relate him in a certain way to honest men. But the relations described in this and in our other psychological statements, if they can properly be called 'relations,' are of a peculiar sort. They can hold even though one of their terms, if it can properly be called a 'term,' does not exist. It may seem, therefore, that one can be 'intentionally related' to something which does not exist.[2]

These points can be put somewhat more precisely by referring to the language we have used. We may say that, in our language, the expressions 'looks for,' 'expects,' and 'believes' occur in sentences which are intentional, or are used intentionally, whereas 'sits in,' 'eats,' and 'shoots' do not. We can formulate a working criterion by means of which we can distinguish sentences that are intentional, or are used intentionally, in a certain language from sentences that are not. It is easy to see, I think, what this criterion would be like, if stated for ordinary English.

First, let us say that a simple declarative sentence is intentional if it uses a substantival expression—a name or a description—in such a way that neither the sentence nor its contradictory implies either that there is or that there isn't anything to which the substantival expression truly applies. 'Diogenes looked for an honest man' is intentional by this criterion. Neither 'Diogenes looked for an honest man' nor its contradictory—'Diogenes did *not* look for an honest man'—implies either that there are, or that there are not, any honest men. But 'Diogenes sits in his tub' is not intentional by this criterion, for it implies that there *is* a tub in which he sits.

Secondly, let us say, of any noncompound sentence which contains a propositional clause, that it is intentional provided that neither the sentence nor its contradictory implies either that the propositional clause is true or that it is false. 'James believes there are tigers in India' is intentional by this criterion, because neither it nor its contradictory implies either that there are, or that there are not, any tigers in India. 'He succeeded in visiting India,' since it implies that he did visit India, is not intentional. 'He is able to visit India,' although it does not imply that he will visit India, is also not intentional. For its contradictory—'He is not able to visit India'—implies that he does *not* visit India.

A third mark of intentionality may be described in this way. Suppose there are two names or descriptions which designate the same things and that *E* is a sentence obtained merely by separating these two names or descriptions by means of 'is identical with' (or 'are identical with' if the first word is plural). Suppose also that *A* is a sentence using one of those names or descriptions and that *B* is like *A* except that, where *A* uses the one, *B* uses the other. Let us say that *A* is intentional if the conjunction of *A* and *E* does not imply *B*.[3] We can now say of certain cognitive sentences—sentences using 'know,' 'see,' 'perceive,' and the like in one of the ways which have interested us here—that they, too, are intentional. Most of us knew in 1944 that Eisenhower was the one in command (*A*); but although he was (identical with) the man who was to succeed Truman (*E*), it is not true that we knew in 1944 that the man who was to succeed Truman was the one in command (*B*).

Let us say that a *compound* sentence is one compounded from two or more sentences by means of propositional connectives, such as 'and,' 'or,' 'if-then,' 'although,' 'because,' and the like. The three foregoing marks of intentionality apply to sentences which are *not* compound. We may now say that a compound declarative sentence is intentional if and only if one or more of its component sentences is intentional. Thus the antecedent of 'If Parsifal sought the Holy Grail, he was a Christian' enables us to say that the whole statement is intentional.

When we use perception words propositionally, our sentences display the third of the above marks of intentionality. I may see that John is the man in the corner and John may be someone who is ill; but I do not now *see* that John is someone who is ill. Perception sentences, as we have seen, entail sentences about taking and assuming. And sentences about taking and assuming display the second of the above marks of intentionality. 'He takes—and therefore assumes—those rocks to be the reef' does not imply that the rocks *are* the reef and it does not imply that they are not. And similarly for its contradiction: 'He does not take—or assume—those rocks to be the reef.'

We may now re-express Brentano's thesis—or a thesis resembling that of Brentano—by reference to intentional sentences. Let us say (1) that we do not need to use intentional sentences when we describe nonpsychological phenomena; we can express all of our beliefs about what is merely 'physical' in sentences which are not intentional.[4] But (2) when we

wish to describe perceiving, assuming, believing, knowing, wanting, hoping, and other such attitudes, then either (a) we must use sentences which are intentional or (b) we must use terms we do not need to use when we describe nonpsychological phenomena.

In describing nonpsychological phenomena, we do, on occasion, use sentences which are intentional by one or more of the above criteria. One may say, 'This weapon, suitably placed, is capable of causing the destruction of Boston' and 'The cash register knows that 7 and 5 are 12.' But although these sentences are intentional according to our criteria, we can readily transform them into others which are not: 'If this weapon were suitably placed, then Boston would be destroyed' and 'If you press the key marked '7' and the one marked '5,' the cash register will yield a slip marked '12."

It would be an easy matter, of course, to invent a psychological terminology enabling us to describe perceiving, taking, and assuming in sentences which are not intentional. Instead of saying, for example, that a man *takes* something to be a deer, we could say 'His perceptual environment is deer-inclusive.' But in so doing, we are using technical terms—'perceptual environment' and 'deer-inclusive'—which, presumably, are not needed for the description of nonpsychological phenomena. And unless we can re-express the deer-sentence once again, this time as a nonintentional sentence containing no such technical terms, what we say about the man and the deer will conform to our present version of Brentano's thesis.

How would we go about showing that Brentano was wrong? I shall consider the three most likely methods. None of them seems to be satisfactory.

3

Some philosophers have tried to describe psychological attitudes in terms of *linguistic* behavior. In his inaugural lecture, *Thinking and Meaning*, Professor Ayer tried to define the locution 'thinking of x' by reference to the use of symbols which designate x. A man is *thinking of* a unicorn, Ayer suggested, if (among other things) the man is disposed to use symbols which *designate* unicorns; he *believes* that there are unicorns if (among other things) he is disposed to utter sentences containing words which *designate* or *refer* to unicorns.[5] And perhaps one might try to define 'taking' and 'assuming' in a similar way. But this type of definition leaves us with our problem.

When we talk about what is 'designated' or 'referred to' by words or sentences, our own sentences are intentional. When we affirm the sentence 'In German, *Einhorn* designates, or refers to, unicorns,' we do not imply that there are any unicorns and we do not imply that there are not; and similarly when we deny the sentence. If we think of words and sentences as classes of noises and marks, then we may say that words and sentences are 'physical' (nonpsychological) phenomena. But we must not suppose the meaning of words and sentences to be a property which they have apart from their relations to the psychological attitudes of the people who *use* them.

For we know, as Schlick once put it, 'that meaning does not inhere in a sentence where it might be discovered'; meaning 'must be bestowed upon' the sentence.[6] Instead of saying, 'In German, *Einhorn* designates, or refers to, unicorns,' we could say, less misleadingly, 'German-speaking people use the word *Einhorn* in order to designate, or refer to, unicorns.' A word or sentence designates so-and-so only if people *use* it to designate so-and-so.

Or can we describe 'linguistic behavior' by means of sentences which are not intentional? Can we define such locutions as 'the word '*Q*' designates so-and-so' in language which is not intentional? If we can do these things, and if, as Ayer suggested, we can define 'believing,' or 'assuming,' in terms of linguistic behavior, then we must reject our version of Brentano's thesis. But I do not believe that we can do these things; I do not believe that we can define such locutions as 'The word '*Q*' designates so-and-so' or 'The word '*Q*' has such-and-such a *use*' in language which is not intentional.

Let us consider, briefly, the difficulties involved in one attempt to formulate such a definition.

Instead of saying, of a certain word or predicate of '*Q*,' that it designates or refers to so-and-so's, we may say that, if there were any so-and-so's, they would satisfy or fulfill the *intension* of the predicate '*Q*.' But how are we to define 'intension'? Professor Carnap once proposed a behavioristic definition of this use of 'intension' which, if it were adequate, might enable us to formulate a behavioristic, nonintentional definition of 'believe' and 'assume.' Although Carnap later conceded that his account was oversimplified, it is instructive, I think, to note the difficulties which stand in the

way of defining 'intension'—as well as 'designates' and 'refers to'—in nonintentional terms.[7]

Carnap had suggested that the 'intension' of a predicate in a natural language may be defined in essentially this way: 'The intension of a predicate '*Q*' for a speaker X is the general condition which an object *y* must fulfill in order for X to be willing to ascribe the predicate '*Q*' to *y*.' Carnap did not define the term 'ascribe' which appears in this definition, but from his general discussion we can see, I think, that he would have said something very much like this: 'A person X ascribes '*Q*' to an object *y*, provided that, in the presence of *y*, X gives an affirmative response to the question '*Q*?'' '(Let us assume that the expressions 'is willing to,' 'in the presence of,' 'affirmative response,' and 'question' present no difficulties.)

Such a definition of 'intension' is adequate only if it allows us to say of Karl, who speaks German, that an object *y* fulfills the intension of '*Hund*' for Karl if and only if *y* is a dog. Let us consider, then, a situation in which Karl mistakes something for a dog; he is in the presence of a fox, say, and takes it to be a dog. In this case, Karl would be willing to give an affirmative response to the question '*Hund*?' Hence the fox fulfills the condition which an object must fulfill for Karl to be willing to ascribe '*Hund*' to it. And therefore the definition is inadequate.

Perhaps we can assume that Karl is usually right when he takes something to be a dog. And perhaps, therefore, we can say this: 'The intension of '*Hund*' for Karl is the general condition which, more often than not, an object *y* must fulfill in order for Karl to be willing to ascribe "*Hund*' to *y*.' But if the occasion we have considered is the only one on which Karl has been in the presence of a *fox*, then, according to the present suggestion, we must say, falsely, that the fox does not fulfill the intension of Karl's word '*Fuchs*.' Moreover, if Karl believes there are unicorns and, on the sole occasion when he thinks he sees one, mistakes a horse for a unicorn, then the present suggestion would require us to say, falsely, that the horse fulfills the intension, for Karl, of his word '*Einhorn*.'

The obvious way to qualify Carnap's definition would be to reintroduce the term 'believe' and say something of this sort: 'The intension of a predicate '*Q*' for a speaker X is the general condition which X must *believe* an object *y* to fulfill in order for X to be willing to ascribe the predicate "*Q*' to *y*.' And, in general, when we say, 'People use such and such a word to refer to so-and-so,' at least part of what we mean to say is that people use that word when they wish to express or convey something they *know* or *believe*—or *perceive* or *take*—with respect to so-and-so. But if we define 'intension' and 'designates' in terms of 'believe' and 'assume,' we can no longer hope, of course, to define 'believe' and 'assume' in terms of 'intension' or 'designates.'

4

The second way in which we might try to show that Brentano was wrong may be described by reference to a familiar conception of 'sign behavior.' Many philosophers and psychologists have suggested, in effect, that a man may be said to *perceive* an object *x*, or to *take* some object *x* to have a certain property *f*, provided only that there is something which *signifies x* to him, or which signifies to him that *x* is *f*. But what does 'signify' mean?

We cannot be satisfied with the traditional descriptions of 'sign behavior,' for these, almost invariably, define such terms as 'sign' by means of intentional concepts. We cannot say, for instance, that an object is a sign provided it causes someone to *believe*, or *expect*, or *think of* something; for sentences using 'believe,' 'expect,' and 'think of' are clearly intentional. Nor can we say merely that an object is a sign provided it causes someone to be *set for*, or to be *ready for*, or to *behave appropriately to* something, for sentences using 'set for,' 'ready for,' and 'behave appropriately to,' despite their behavioristic overtones, are also intentional. Similar objections apply to such statements as 'One object is a sign of another provided it *introduces* the other object *into the behaviorial environment*, as contrasted with the physical environment, of some organism.'

If we are to show that Brentano's thesis as applied to *sign* phenomena is mistaken, then we must not introduce any new technical terms into our analysis of sign behavior unless we can show that these terms apply also to nonpsychological situations.

Most attempts at nonintentional definitions of 'sign' make use of the concept of *substitute stimulus*. If we use 'referent' as short for 'what is signified,' we may say that, according to such definitions, the sign is described as a substitute for the referent. It is a substitute in the sense that, as stimulus, it has effects upon the subject which are similar to those the referent would have had. Such definitions usually take

this form: *V* is a *sign* of *R* for a subject *S* if and only if *V* affects *S* in a manner similar to that in which *R* would have affected S.[8] The bell is a sign of food to the dog, because the bell affects the dog's responses, or his dispositions to respond, in a way similar to that in which the food would have affected them.

This type of definition involves numerous difficulties of which we need mention but one—that of specifying the respect or degree of similarity which must obtain between the effects attributed to the sign and those attributed to the referent. This difficulty is involved in every version of the substitute-stimulus theory. Shall we say that, given the conditions in the above definition, *V* is a sign of *R* to a subject *S* provided only that those responses of *S* which are stimulated by *V* are similar in *some* respect to those which have been (or would be) stimulated by R? In other words, should we say that *V* is a sign of *R* provided that *V* has some of the effects which *R* has had or would have had? This would have the unacceptable consequence that all stimuli signify each other, since any two stimuli have at least some effect in common. Every stimulus causes neural activity, for example; hence, to that extent at least, any two stimuli will have similar effects. Shall we say that *V* is a sign of *R* provided that *V* has *all* the effects which *R* would have had? If the bell is to have all the effects which the food would have had, then, as Morris notes, the dog must start to eat the bell.[9] Shall we say that *V* is a sign of *R* provided that *V* has the effects which *only R* would have had? If the sign has effects which only the referent can have, then the sign *is* the referent and only food can be a sign of food. The other methods of specifying the degree or respect of similarity required by the substitute-stimulus definition, so far as I can see, have equally unacceptable consequences.

Reichenbach, in his *Elements of Symbolic Logic*, has applied this type of analysis to the concept of taking; but the consequences are similar. To say of a subject S, according to Reichenbach, that *S takes* something to be a dog is to say: 'There is a *z* which is a bodily state of *S* and which is such that, whenever *S* is sensibly stimulated by a dog, *S* is in this bodily state *z*.'[10] In other words, there are certain bodily conditions which *S* must fulfill in order for *S* to be sensibly stimulated by a dog; and whenever *S* satisfies any of these conditions, then *S* is taking something to be a dog.

But among the many conditions one must fulfill if one is to be sensibly stimulated by a dog is that of being alive. Hence if we know that *S* is alive, we can say that *S* is taking something to be a dog. The difficulty is that the bodily state *z*, of Reichenbach's formula, is not specified strictly enough. And the problem is to find an acceptable modification.

In reply to this objection, Reichenbach suggested, in effect, that '*S* takes something to be a dog' means that *S*'s bodily state has all those neural properties which it must have—which are 'physically necessary' for it to have—whenever *S* is sensibly stimulated by a dog.[11] But this definition has the unacceptable consequence that, whenever *S* is sensibly stimulated by a dog, then *S takes* the thing to be a dog. Thus, although we can say that a man may be stimulated by a fox and yet take it to be a dog, we can never say that he may be stimulated by a dog and *not* take it to be a dog.[12]

Similar objections apply to definitions using such expressions as 'dog responses,' 'responses specific to dogs,' 'responses appropriate to dogs,' and the like. For the problem of specifying what a man's 'dog responses' might be is essentially that of specifying the bodily state to which Reichenbach referred.

5

Of all intentional phenomena, expectation is one of the most simple and, I think, one which is most likely to be definable in terms which are not intentional. If we could define, in nonintentional terms, what it means to say of a man, or an animal, that he expects something—that he expects some state of affairs to come about—then, perhaps, we could define 'believing' and 'assuming,' nonintentionally, in terms of this sense of 'expecting.' If we are to show that Brentano is wrong, our hope lies here, I think.

For every expectancy, there is some possible state of affairs which would *fulfill* or *satisfy* it, and another possible state of affairs which would *frustrate* or *disrupt* it. If I expect the car to stop, then, it would seem, I am in a state which would be fulfilled or satisfied if and only if the car were to stop—and which would be frustrated or disrupted if and only if the car were not to stop. Hence we might consider defining 'expects' in this way:

> '*S expects E* to occur' means that *S* is in a bodily state *b* such that either (i) *b* would be fulfilled if and only if *E* were to occur or (ii) *b* would be disrupted if and only if *E* were not to occur.

Our problem now becomes that of finding appropriate meanings for 'fulfill' and 'disrupt.'

Perhaps there is a way of defining 'fulfill' in terms of the psychological concept of *reinforcement* and of defining 'disrupt' in terms of *disequilibration, surprise*, or *shock*. And perhaps we can then provide an account of the dog and the bell and the food in terms which will show that this elementary situation is not intentional. It is possible that the dog, because of the sound of the bell, is in a state which is such that either (i) his state will be reinforced if he receives food or (ii) it will be disequilibrated if he does not. And it is possible that this state can be specified in physiological terms. Whether this is so, of course, is a psychological question which no one, apparently, is yet in a position to answer. But even if it is so, there are difficulties in principle which appear when we try to apply this type of definition to human behavior.

If we apply 'expects,' as defined, to human behavior, then we must say that the appropriate fulfillments or disruptions must be caused by the occurrence, or nonoccurrence, of the 'intentional object'—of *what* it is that is expected. But it is easy to think of situations which, antecedently, we should want to describe as instances of expectation, but in which the fulfillments or disruptions do not occur in the manner required. And to accommodate our definition to such cases, we must make qualifications which can be expressed only by reintroducing the intentional concepts we are trying to eliminate.

This difficulty may be illustrated as follows: Jones, let us suppose, *expects* to meet his aunt at the railroad station within twenty-five minutes. Our formulation, as applied to this situation, would yield: 'Jones is in a bodily state which would be fulfilled if he were to meet his aunt at the station within twenty-five minutes or which would be disrupted if he were not to meet her there within that time.' But what if he were to meet his aunt and yet *take* her to be someone else? Or if he were to meet someone else and yet *take* her to be his aunt? In such cases, the fulfillments and disruptions would not occur in the manner required by our definition.

If we introduce the intentional term 'perceives' or 'takes' into our definition of 'expects,' in order to say, in this instance, that Jones *perceives* his aunt, or *takes* someone to be his aunt, then, of course, we can no longer define 'assume'—or 'perceive' and 'take'—in terms of 'expects.' It is worth noting, moreover, that even if we allow ourselves the intentional term 'perceive' our definition will be inadequate. Suppose that Jones were to visit the bus terminal, believing it to be the railroad station, or that he were to visit the railroad station believing it to be the bus terminal. If he met his aunt at the railroad station, believing it to be the bus terminal, then, contrary to our formula, he may be frustrated or surprised, and, if he fails to meet her there, his state may be fulfilled. Hence we must add further qualifications about what he believes or doesn't believe.[13]

If his visit to the station is brief and if he is not concerned about his aunt, the requisite reinforcement or frustration may still fail to occur. Shall we add '. . . provided he *looks for* his aunt'? But now we have an intentional expression again. And even if we allow him to look for her, the re-enforcement or frustration may fail to occur if he finds himself able to satisfy desires which are more compelling than that of finding his aunt.

We seem to be led back, then, to the intentional language with which we began. In attempting to apply our definition of 'expects' to a situation in which 'expects' is ordinarily applicable, we find that we must make certain qualifications and that these qualifications can be formulated only by using intentional terms. We have had to introduce qualifications wherein we speak of the subject *perceiving* or *taking* something to be the object expected; hence we cannot now define 'perceive' and 'assume' in terms of 'expect.' We have had to add that the subject has certain *beliefs* concerning the nature of the conditions under which he perceives, or fails to perceive, the object. And we have referred to what he is *looking for* and to his other possible *desires*.

It may be that some of the simple 'expectancies' we attribute to infants or to animals can be described, nonintentionally, in terms of reinforcement or frustration. And possibly, as Ogden and Richards intimated, someone may yet find a way of showing that believing, perceiving, and taking are somehow 'theoretically analysable' into such expectancies.[14] But until such programs are carried out, there is, I believe, some justification for saying that Brentano's thesis does apply to the concept of *perceiving*.

NOTES

1. Franz Brentano, *Psychologie vom empirischen Standpunkte* (Leipzig: Felix Meiner, 1924), 1, pp. 24–25.
2. But the point of talking about 'intentionality' is not that there is a peculiar type of 'inexistent' object; it is rather that there is a type of psychological phenomenon which is unlike anything purely physical. In his later writings Brentano explicitly rejected the view that there are 'inexistent objects'; see his *Psychologie*, II, pp. 133 ff., and *Wahrheit und Evidenz* (Leipzig: Felix Meiner, 1930), pp. 87, 89.
3. This third mark is essentially the same as Frege's concept of 'indirect reference.' See Gottlob Frege, "Über Sinn und Bedeutung," *Zeitschrift für Philosophie und philosophische Kritik*, n.s. C (Halle: Pfeffer and Leipzig: Haacke, 1892), pp. 25–50, especially 38; reprinted in Herbert Feigl and W. S. Sellars, eds., *Readings in Philosophical Analysis* (New York: Appleton, 1949), and Peter Geach and Max Black, eds., *Philosophical Writings of Gottlob Frege* (Oxford: Blackwell, 1952).
4. There are sentences describing relations of comparison—for example, 'Some lizards look like dragons'—which may constitute an exception to (1). If they are exceptions, then we may qualify (1) to read: 'We do not need any intentional sentences, other than those describing relations of comparison, when we describe nonpsychological phenomena.' This qualification would not affect any of the points to be made here.
5. A. J. Ayer, *Thinking and Meaning*, (London, Lewis, 1947), p. 13. Compare W. S. Sellars, "Mind, Meaning, and Behavior," *Philosophical Studies* III (1952): pp. 83–95; "A Semantical Solution of the Mind-Body Problem," *Methodos* (1953); pp. 45–85; and "Empiricism and the Philosophy of Mind," in *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, Herbert Feigl and Michael Scriven, eds., (Minneapolis: University of Minnesota Press, 1956). See also Leonard Bloomfield, *Linguistic Aspects of Science* (Chicago: University of Chicago Press, 1939), pp. 17–19.
6. Moritz Schlick, "Meaning and Verification," *Philosophical Review* XLV (1936): p. 348; reprinted in Herbert Feigl and W. S. Sellars, eds., *Readings in Philosophical Analysis.* Compare this analogy in "Meaning and Free Will," by John Hospers: 'Sentences in themselves do not possess meaning; it is misleading to speak of 'the meaning of sentences' at all; meaning being conferred in every case by the speaker, the sentence's meaning is only like the light of the moon: without the sun to give it light, it would possess none. And for an analysis of the light we must go to the sun' (*Philosophy and Phenomenological Research*, X [1950]: p. 308).
7. Carnap's definition appeared on p. 42 of "Meaning and Synonymy in Natural Languages," *Philosophical Studies*, IV (1955)" pp. 33–47. In "On Some Concepts of Pragmatics," *Philosophical Studies*, VI : pp. 89–91, he conceded that 'designates' should be defined in terms of 'believes.' The second article was written in reply to my "A Note on Carnap's Meaning Analysis," which appeared in the same issue (pp. 87–89).
8. Compare Charles E. Osgood, *Method and Theory in Experimental Psychology* (New York: Oxford University Press, 1953), p. 696: 'A pattern of stimulation which is not the object is a sign of the object if it evokes in an organism a mediating reaction, this (a) being some fractional part of the total behavior elicited by the object and (b) producing distinctive self-stimulation that mediates responses which would not occur without the previous association of nonobject and object patterns of stimulation. All of these limiting conditions seem necessary. The mediation process must include part of the same behavior made to the object if the sign is to have its representing property.' Some of the difficulties of the substitute stimulus concept [qualification (a) in this definition] are met by qualification (b), which implies that the subject must once have perceived the thing signified. But (b) introduces new difficulties. Since I have never seen the President of the United States, no announcement, according to this definition, could signify to me that the President is about to arrive.
9. See Charles Morris, *Signs, Language, and Behavior* (New York, Oxford University Press, 1953), p. 12, and Max Black, "The Limitations of a Behavioristic Semiotic," *Philosophical Review* LVI (1947): pp. 258–72.
10. This is a paraphrase of what Hans Reichenbach formulated in special symbols on p. 275 of *Elements of Symbolic Logic* (New York: Macmillan Co., 1947).
11. Reichenbach suggests this modification in "On Observing and Perceiving," *Philosophical Studies* II (1951): pp. 92–93. This paper was written in reply to my "Reichenbach on Observing and Perceiving" (*Philosophical Studies* II (1951): pp. 45–48), which contains some of the above criticisms. In these papers, as well as in Reichenbach's original discussion, the word 'perceive' was used in the way in which we have been using 'take.' Reichenbach used the term 'immediate existence' in place of Brentano's 'intentional inexistence'; see *Elements of Symbolic Logic*, p. 274.
12. This sort of modification may suggest itself: Consider those bodily states which are such that (i) *S* is in those states whenever he is sensibly stimulated by a dog and (ii) *S* cannot be in those states whenever he is *not* being stimulated by a dog. Shall we say '*S* takes something to be a dog' means that *S* is in this particular class of states? If we define 'taking' in this way, then, we must say that, in the present state of psychology and physiology, we have no way of knowing whether anyone ever *does* take anything to be a dog, much less whether people take things to be dogs on just those occasions on which we want to be able to *say* that they take things to be dogs.
13. R. B. Braithwaite in "Belief and Action" (*Aristotelian Society*, suppl. vol. XX [1946]: p. 10) suggests that a man may be said to believe a proposition *p* provided this condition obtains: 'If at a time when an occasion arises relevant to *p*, his springs of action are *s*, he will perform an action which is such that, if *p* is true, it will tend to fulfill *s*, and which is

such that, if p is false, it will not tend to satisfy s.' But the definition needs qualifications in order to exclude those people who, believing truly (p) that the water is deep at the base of Niagara Falls and wishing (s) to survive a trip over the falls, have yet acted in a way which has not tended to satisfy s. Moreover, if we are to use such a definition to show that Brentano was wrong, we must provide a non-intentional definition of the present use of 'wish' or 'spring of action.' And, with Braithwaite's definition of 'believe,' it would be difficult to preserve the distinction which, apparently, we ought to make between *believing* a proposition and *acting upon* it (see chapter one, section 2). I have proposed detailed criticisms of a number of such definitions of 'believe' in "Sentences about Believing," *Proceedings of the Aristotelian Society* LVI (1955–1956): pp. 125–48. Some of the difficulties involved in defining *purpose* nonintentionally are pointed out by Richard Taylor in "Comments on a Mechanistic Conception of Purpose," *Philosophy of Science* XVII (1950): pp. 310–17, and "Purposeful and Nonpurposeful Behavior: A Rejoinder," ibid., pp. 327–32.

14. C. K. Ogden and I. A. Richards, *The Meaning of Meaning* (London: 1938), 5th ed., p. 71.

A Recipe for Thought

Fred Dretske

1. If You Can't Make One, You Don't Know How It Works

There are things I believe that I don't know how to say—at least not in such a way as to make them come out true. The title of this section is a case in point. I really do believe that, in the relevant sense of all the relevant words, if you can't make one, you don't know how it works. I just don't know how to specify the relevant sense of all the relevant words.

I know, for instance, that a person can understand how something works and, for a variety of reasons, still not be able to build it. The raw materials are not available. She can't afford them. He is too clumsy or not strong enough. The police won't let him. I also know that a person may be able to make one and still not know how it works. He doesn't know how the parts work. I can solder a snaggle to a radzak, and this is all it takes to make a gizmo, but if I do not know what snaggles and radzaks are, or how they work, making one isn't going to tell me much about what a gizmo is. My son once assembled a television set from a kit by carefully following the instruction manual. Understanding next to nothing about electricity, though, he still had no idea of how television worked.

I am not, however, suggesting that being able to build one is sufficient for knowing how it works. Only necessary. And I do not much care about whether you can actually put one together. It is enough if you know how one is put together. But, as I said, I do not know how to make all the right qualifications. So I won't try. All I intend by my provocative claim is that philosophical naturalism is motivated by a constructivist model of understanding. It embodies something like an engineer's ideal, a designer's vision, of what it takes to really understand how something works. You need a blueprint, a recipe, an instruction manual, a program. That goes for the mind as well as anything else. If you want to know what intelligence is, you need a recipe for creating it out of parts you already understand.

In speaking of parts one already understands, I mean, of course, parts that do not already possess the capacity or feature one follows the recipe to create. One cannot have a recipe for a cake that lists a cake, not even a small cake, as an ingredient. One can, I suppose, make a big cake out of small cakes, but recipes of this sort will not help one understand what a cake is (though they might help you understand what a *big* cake is). As a boy, I once tried to make

Originally published as "If You Can't Make One, You Don't Know How It Works," in P. French, T. Uehling, and H. Wettstein, eds., *Midwest Studies in Philosophy*, vol. 19 (1994), pp. 468–82. Reprinted (with small revisions by the author) with the permission of the publisher.

fudge by melting caramels in a frying pan. All I succeeded in doing was ruining the pan. Don't ask me what I was trying to do—change the shape of the candy, I suppose. There are perfectly respectable recipes for cookies that list candy (e.g., gumdrops) as an ingredient, but one cannot have a recipe for *candy* that lists candy as an ingredient. At least it won't be a recipe that tells you how to make candy or helps you understand what candy is. The same is true of minds. That is why recipes for thought can't have interpretive attitudes or explanatory stances among the ingredients—not even the attitudes and stances of others. That is like making candy out of candy—in this case, one person's fudge out of another person's caramels. You can do it, but you still won't know what candy is.

2. Information and Intentionality

In comparing a mind to candy and television, I do not mean to suggest that minds are the sort of thing that can be assembled from kits in your basement or kitchen. There are things, including things one fully understands, things one knows how to make, that cannot be assembled that way. Try making Rembrandts or one hundred dollar bills in your basement. What you produce may look genuine, it may pass as authentic, but it won't be the real thing. You have to be the right person, occupy the right office, or possess the appropriate legal authority in order to make certain objects. There are recipes for making money and Rembrandts, and knowing them is part of knowing what money and Rembrandts are, but these are not recipes you and I can use. Some recipes require a special cook.

This is one (but only one) of the reasons it is wrong to say, as I did above, that if you cannot make one, you do not know how it works. It would be better to say, as I did earlier, that if you do not know *how* to make one, or *how* one is made, you do not fully understand it.

Some objects are constituted, in part, by their relationships to other objects. Rembrandts and one hundred dollar bills are like that. So are cousins and mothers-in-law. That is why you can't build my cousin in your basement while my aunt and uncle can. Though there is a recipe knowledge of which is necessary for understanding what it takes to be my cousin, it is not a recipe *you* can use to build what it enables you to understand. The mind, I think, is like that, and I will return to this important point in a moment.

It is customary to think of naturalistic recipes for the mind as starting with extensional ingredients and, through some magical blending process, producing an intentional product: a thought, an experience, or a purpose. The idea behind this proscription of intentional ingredients seems to be that since what we are trying to build—a thought—is an intentional entity, our recipe cannot use intentional ingredients.

This, it seems to me, is a mistake, a mistake that has led otherwise sensible philosophers to despair of ever finding a naturalistic recipe for the mind. It is a mistake that has given naturalism an undeserved bad name. The mistake is the same as if we proscribed using, say, copper wire in our instruction manual for building an amplifier because copper wire conducts electricity—exactly what the thing we are trying to build—an amplifier—does. But there is nothing wrong in listing copper wire in one's recipe for building an amplifier. An amplifier recipe is supposed to help you make (and, thus, understand) how things amplify electricity, not how something conducts electricity. That is why you get to use conductors of electricity as components in a recipe for building an amplifier. Conductors are eligible ingredients in amplifier recipes even if one does not know how conductors manage to conduct. An eligible ingredient, once again, is an ingredient, a part, a component, that does not already have the capacity or power one follows the recipe to create. That is why one can know what gumdrop cookies are, know how to make them, without knowing how to make gumdrops or what, exactly, gumdrops are.

The same is true for mental recipes. As long as there is no mystery—at least not the *same* mystery—about how the parts work as how the whole is supposed to work, it is perfectly acceptable to use intentional ingredients in a recipe for thought, purpose, and intelligence. What we are trying to understand, after all, is not intentionality, per se, but the mind. Thought may be intentional, but that isn't the property we are seeking a recipe to understand. As long as the intentionality we use is not itself mental, then we are as free to use intentionality in our recipe for making a mind as we are in using electrical conductors in building an amplifier or gumdrops in making cookies.

Consider a simple artifact—a compass. If it was manufactured properly (don't buy a cheap

one), and if it is used in the correct circumstances (the good ones come with directions), it will tell you the direction of the arctic pole.[1] That is what the pointer indicates. But though the pointer indicates the direction of the arctic pole, it does not indicate the whereabouts of polar bears even though polar bears live in the arctic. If you happen to know this fact about polar bears, that they live in the arctic, you could, of course, figure out where the polar bears are by using a compass. But this fact about what you could figure out *if you knew* does not mean that the compass pointer is sensitive to the location of polar bears—thus indicating their whereabouts—in the way it indicates the location of the arctic pole. The pointer on this instrument does not track the bears; it tracks the pole. If there is any doubt about this, watch the compass needle as you move the polar bears around. It won't even wiggle.

Talking about what a compass indicates is a way of talking about what it tracks, what information it carries, and a compass, just like any other measuring instrument, can track one magnitude without tracking another even though these conditions co-occur. Talk about what instruments and gauges indicate or measure creates the same kind of intensional (with an 's') context as does talk about what a person knows or believes. Knowing or believing that *that* is the north pole is not the same as knowing or believing that that is the habitat of polar bears even though the north pole is the habitat of polar bears. If we use intensional (with an 's') discourse, referentially opaque contexts, as a guide to intentional (with a 't') phenomena, then we have, in a cheap compass, something we can buy at the local hardware store, intentionality. Describing what such an instrument indicates is describing it in intensional terms. What one is describing with these intensional terms is, therefore, in this sense, an intentional state of the instrument.

It is worth emphasizing that this is not derived or in any way second-class intentionality. This is the genuine article—*original* intentionality as some philosophers (including this one) like to say. The intentional states a compass occupies do not depend on our explanatory purposes, attitudes, or stances. To say that the compass indicates the direction of the arctic pole is to say that the position of the pointer depends on the whereabouts of the pole. This dependency exists whether or not we know it exists, whether or not anyone ever exploits this fact to build and use compasses. The intentionality of the device is not like the intentionality of words and maps, borrowed or derived from the intentionality (purposes, attitudes, knowledge) of its users. The power of this instrument to indicate north *to* or *for us* may depend on our taking it to be a reliable indicator (and, thus, on what we believe or know about it), but its *being* a reliable indicator does not depend on us.

Intentionality is a much-abused word and it means a variety of different things. But one thing it has been used to mean is some state, condition, activity or event, whose description generates an opaque context, a context in which coextensional terms cannot be automatically substituted for one another. This is what Chisholm describes as the third mark of intentionality.[2] Anything exhibiting this mark is about something under an aspect. It has an aspectual shape.[3] The compass needle is about the arctic under one aspect (as the location of the north pole) and not others (as the habitat of polar bears). This is the same way our thoughts can be about a place under one aspect (as where I was born) but not another (as where you were born). If this is, indeed, one thing that is meant by speaking of a state, condition, or activity as intentional, then it seems clear that there is no need to naturalize intentionality. It is already a completely natural phenomenon, a pervasive feature of our physical world. It exists wherever you find dark clouds, smoke, tree rings, shadows, tracks, lightning, flowing water, and countless other natural conditions that indicate something about how the rest of the world is constituted.

Intentional systems, then, are not the problem. They can be picked up for a few dollars at your local hardware store. We can, therefore, include them on our list of ingredients in our recipe for building a mind without fear that we are merely changing the shape of the candy. What we are trying to build when we speak of a recipe for building a mind is not merely a system that exhibits intentional properties. We already have that in systems that are in no way mental. Rather, what we are trying to build is a system that exhibits that peculiar array of intentional properties that characterizes thought. We are, in particular, trying to build systems that exhibit what Chisholm describes as the first mark of intentionality, the power to say that so-and-so is the case when so-and-so is not the case, the power to misrepresent how things stand in the world. Unlike compasses, these fancy items are not to be found on the shelves of hardware stores. For them we need a recipe.

3. Misrepresentation

Let us be clear about what we are looking for, what we seek a recipe to create. If we are trying to build a thought, we are looking for something that can not only say that x is F without saying x is G (despite the co-extensionality of 'F' and 'G'[4]), thus being about x under an aspect, we are looking for something that can say this, like a thought can say it, even when x is not F. Without this, we have no naturalistic understanding of what it is we think, no theory of meaning or content. For meaning or content, the what-it-is one thinks, is, like intelligence and rationality, independent of the truth of what one thinks. So a recipe for understanding misrepresentation is, in effect, a recipe for constructing meanings and, therefore, genuinely intelligent systems.

Jerry Fodor has recently focused attention on what he calls the disjunction problem for naturalistic theories of mental representation.[5] The problem is one of explaining how, in broadly causal terms, a structure in the head, call it R, could represent, say, or mean that something was F even though a great many things other than something's being F are capable of causing R. How can the occurrence of R mean that something is F when something's being F is only one of the things capable of causing R?[6] For someone trying to formulate an information-based recipe for thought, this is, indeed, a vexing problem. But I mention the problem here only to point out that this problem is merely another way of describing the problem of misrepresentation. For if one could specify a recipe for building systems capable of misrepresentation—capable, that is, of saying that something was F when it wasn't—then one would have a recipe for meaning, a recipe for constructing structures having a content that was independent of causes. For anything that can misrepresent something as being F is, of necessity, something whose meaning is independent of its causes, something that can mean cow even when it is caused by a horse on a dark night. It is, therefore, something whose meaning is less than the disjunction of conditions capable of causing it, something whose meaning (in the words of Antony and Levine[7]) is 'detached' from causes. A naturalistic recipe for misrepresentation, then, is a recipe for solving the disjunction problem.[8] One way of solving problems is to show that two problems are really, at bottom, the same problem. So we are making progress.

For this problem artifacts are of no help. Although clocks, compasses, thermometers, and fire alarms—all readily available at the corner hardware store—can misrepresent the conditions they are designed to deliver information about, they need our help to do it. Their representational successes and failures are under-written by the purposes and attitudes of their designers and users. As representational devices, as devices exhibiting a causally detached meaning, such instruments are not, therefore, eligible ingredients in a recipe for making thought.

The reason the representational powers of instruments are not, like their indicative (information-carrying) powers, an available ingredient in mental recipes is, I hope, obvious enough. I will, however, take a moment to expand on it in order to set the stage for what follows.

Consider the thermometer. Since the volume of a metal varies lawfully with the temperature, both the mercury in the glass tube and the paper clips in my desk drawer carry information about the local temperature. Both are intentional systems in that minimal, that first, sense already discussed. Their behavior depends on a certain aspect of their environment (on the temperature, not the color or size, of their neighbors) in the same way the orientation of a compass needle depends on one aspect of its environment, not another. The only relevant difference between thermometers and paper clips is that we have given the one volume of metal—the mercury in the glass tube—the job of telling us about temperature. The paper clips have been given a different job. Since it is the thermometer's job to provide information about temperature, it (we say) misrepresents the temperature when it fails to do its assigned job just as (we say) a book or a map might misrepresent the matters about which they purport to inform us. What such artifacts say or mean is what they have the job of indicating, and since you do not lose your job—at least not immediately—merely by failing to successfully perform your job, these instruments continue to mean that a certain condition exists even when something *else* causes them to perform. Meanings are causally detached from causes for the same reason that functions are causally detached from actual functioning. This is why thermometers can, while paper clips cannot, 'say' something false about temperature.

But, as I said, thermometers can't do this by themselves. They need our help. We are the source of the job, the function, without which

the thermometer could not say anything false. Take us away and all you have is a tube full of mercury being caused to expand and contract by changes in the temperature—a column of metal doing exactly what paper clips, thumb tacks, and flag poles do. Once we change our attitude, once we stop investing informational trust in it, the instrument loses its power to misrepresent. Its meaning ceases to be detached. It becomes merely a purveyor of information.

4. Natural Functions

Though representational artifacts are thus not available as eligible ingredients in our recipe for the mind, their derived (from us) power to misrepresent is suggestive. If an information-carrying element in a system could somehow acquire the function of carrying information, and acquire this function in a way that did not depend on our intentions, purposes, and attitudes, then it would thereby acquire (just as a thermometer or a compass acquires) the power to misrepresent the conditions it had the function of informing about. Such functions would bring about a detachment of meaning from cause. Furthermore, since the functions would not be derived from us, the meanings (unlike the meaning of thermometers and compasses) would be original, underived, meaning. Instead of just being able to build an instrument that could fool us, the thing we build could, quite literally, itself be fooled.

If, then, we could find naturalistically acceptable functions, we could combine these with natural indicators (the sort used in the manufacture of compasses, thermometers, pressure gauges, and electric eyes) in a naturalistic recipe for thought. If the word 'thought' sounds a bit fancy for the contraption we are assembling, we can describe the results in more modest terms. What we would have is a naturalistic recipe for representation, a product that would have, quite apart from its creator's (or anyone else's) purposes, attitudes, or thoughts, a propositional content that could be false. If that isn't quite a recipe for Béarnaise sauce, it is at least a recipe for a passable gravy. I'll come back to the Béarnaise sauce in a moment.

What we need in the way of another ingredient, then, is some process whereby elements can acquire, on their own, an information-carrying function. Where might we find these natural processes? There are, as I see it, two possible sources: one phylogenic, the other ontogenic.

If the heart and kidneys have a natural function, something they are supposed to be doing independently of our knowledge or understanding of what it is, then it presumably comes from their evolutionary, their selectional, history.[9] If the heart has the function of pumping blood, if that is why it is there,[10] then, by parity of reasoning, the senses (depending on actual selectional history) might have an information-providing function, the job of 'telling' the animal in whom they occur what it needs to know in order to find food and mates and avoid danger. If this were so, then, the natural function of sensory systems would be to provide information about an organism's optical, acoustic, and chemical surroundings. There would thus exist, inside the animal, representations of its environment, elements capable of saying something false. Though I have put it quite crudely, this, I take it, is the idea that inspires biologically oriented approaches to mental representation.[11]

There is, however, a second, an ontogenetic, source of usable (in naturalistic recipes) functions. Think of a system with needs, certain things it must have in order to survive.[12] In order to satisfy those needs it has to do A in conditions C. Nature has not equipped this system with a mechanism that will automatically trigger A in conditions C. There is, in other words, no instinct to A in circumstances C. Maybe C is a condition that has only recently appeared in this animal's natural habitat. Think of C as an attractive (to this kind of animal) mushroom that is quite poisonous. The animal has the sensory resources for picking up information about (i.e., registering) the presence of C (it looks distinctive), but it does not have an instinctive, a genetically hard-wired, reaction to C. It can perceive C, but it has not yet learned to avoid C. We could wait for natural selection to solve this problem for the species, for the descendants of this animal, but if the problem—basically a coordination problem—is to be solved at the individual level (if *this* animal is to survive)—learning must occur. Some internal sign or indicator of C—the animal's sensory registration of C—must be made into a cause of A. Control circuits must be reconfigured by inserting this internal sign into the behavioral chain of command. Short of a miracle—the fortuitous occurrence of A whenever C is encountered—this is the only way the coordination problem can be solved. The internal indicators must be harnessed to effector mechanisms so as to coordinate output to the conditions they carry information about. Learning of this kind

has the same results for the individual as do the longer-term evolutionary solutions for the species: internal elements that supply needed information acquire the function of supplying it by being drafted into the control loop because they supply it.[13] They are there, doing what they are doing, *because* they supply this information.

Obviously this ingredient, this source of natural functions, cannot be ordered from a spare parts catalog. There is nothing one can squirt on a temperature indicator that will give it the function of indicating temperature, nothing we can rub on photo-sensitive pigment that will give it the job of detecting light. If something is going to get the function, the job, the purpose, of carrying information in this way, it has to get it on its own. We can't give it.[14] If the only natural functions are those provided by evolutionary history and learning, then, no one is going to build a thinker of thoughts, much less a mind, in the laboratory. This would be like building a heart, a real one, in your basement. If hearts are essentially organs of the body having the biological function of pumping blood, you can't build them. You can wait for them to develop, maybe even hurry things along a bit by timely assists, but you can't assemble them out of ready-made parts. These functions are results of the right kind of history, and you cannot, not *now*, give a thing the right kind of history. There is a recipe for building internal representations, but it is not a recipe you or I, or anyone else, can use to build one.

5. The Disjunction Problem

There are reasonable doubts about whether a recipe consisting of information and natural teleology (derived from natural functions—either phylogenic or ontogenic) is capable of yielding a mental product—something with an original power to misrepresent. The doubts exist even with those who share the naturalistic vision. Jerry Fodor, for instance, does not think Darwin (or Skinner, for that matter) can rescue Brentano's chestnuts from the fire.[15] Teleological stories about intentionality, he says, do not solve the disjunction problem. Given the equivalence of the disjunction problem and the problem of misrepresentation, this is a denial, not just a doubt, that evolutionary or learning-theoretic accounts of functions are up to the task of detaching meaning from cause, of making something say COW when it is caused by something other than a cow.[16]

I agree with Fodor about the irrelevance of Darwin for understanding mental representation. I agree, however, not out of a general skepticism about teleological accounts of meaning, but because I think Darwin is the wrong place to look for the functions underlying the kind of *mental* representations (beliefs, thoughts, judgments, preferences, and their ilk) that explain *action*—the sort of voluntary or deliberate behavior for which we typically have reasons. I expect Darwin to help us understand why people blink, reflexively, when someone pokes a finger at their eye, but not why they (deliberately) wink at their friend. There are probably internal representations (of objects approaching the eye) involved in the blink reflex, representations that have an evolutionary origin, but these are not the sort of representations (beliefs, purposes, and intentions) at work in explaining why we wink at a friend or pack for a trip. If we are looking for a naturalized semantics for thought, the sort of representation that helps explain action, we will have to get our teleology from somewhere else. Darwin won't help us because Darwin is concerned with precisely those behaviors the explanatory mechanisms for which are genetically determined—precisely those behaviors that are not voluntary.

Nonetheless, wherever we get the teleology, Fodor thinks it won't solve the disjunction problem and is, therefore, hopeless as an account of thought content. I disagree. I have tried to supply the details in *Explaining Behavior* so I won't repeat myself here. Let me here mention only a crucial point. An historical theory of content is not, as Fodor thinks, restricted to assigning content in terms of the objects or conditions that actually figured in the development of the representation. If R, a COW indicator, gets its function of indicating cows by 'exposure' to only Jersey cows, this does not mean that R means (has as its representational content) JERSEY COW. Whether it means COW, JERSEY COW, or, perhaps, simply ANIMAL, will depend (as Fodor likes to say) on the counterfactuals. In indicating that yonder object is a Jersey cow, R also indicates that it is a cow and, therefore, an animal. It indicates all these things. But though R carries all these pieces of information, a developmental theory of content identifies what R has the function of indicating—hence, what R represents—with that particular piece of information that was causally relevant in the selectional process by means of which R was recruited for causal duties. Was it JERSEY COW, COW,

or, perhaps, simply, ANIMAL, the indication of which led to R's recruitment as a determinant of system output? The answer to this question is an answer to the question, 'What does R represent?' and it requires an evaluation of the counterfactuals that Fodor thinks relevant to a determination of content.

6. The Recipe

What we have, then, is the following recipe for making a thought. It does not give us a very fancy thought—certainly nothing like the thoughts we have every day: that tomorrow is my birthday or that I left my umbrella in the car. But one thing at a time. The recipe will do its job if it yields *something*—call it a 'protothought'—that has belief-like features. We can worry about the fancy trimmings later.

RECIPE FOR THOUGHT: Take a system that has a need (see footnote 12) for the information that F, a system whose survival or well-being depends on its doing A in conditions F. Make sure that this system has a means of detecting (i.e., an internal element that indicates) the presence of condition F. Add a natural process, one capable of conferring on the element that carries information F the function of carrying this piece of information. Of course, you don't just 'add' this process the way you add spices in a recipe for lasagna. Adding the function is more like *waiting* for dough to rise. There is nothing you can do but sit back and hope things develop in the right way. And just as you cannot put yeast in just anything and expect it to rise (it doesn't work in sand), you cannot put indicators of F in just anything having a need for this information and expect it to spontaneously generate representations of F. You need a system with the capacity to reorganize control circuits so as to exploit this information in achieving coordination of its behavior with the conditions (F) it is getting information about. These are pretty special sorts of systems, to be sure. They are systems capable of learning. I have no doubt that living systems of a certain level of complexity are the only ones able to perform the trick. However special they might be, though, they needn't be systems that already possess powers of representation. In requiring systems of this sort, therefore, we are not using tainted ingredients in our recipe for thought.

If all goes well, when the process is complete, the result will be a system with internal resources for representing (with the associated power of misrepresenting) its surroundings. Furthermore, that this system represents, as well as what it represents, will be independent of what we know or believe about it. For we, the cooks, are not essential parts of this process. The entire process can happen spontaneously and, when it does, the system will have its own cache of *original* intentionality.

7. Rationality: The Functional Role of Thought

Whether this is really enough to have supplied a recipe for thought depends, of course, on just what one demands of thought. What does it take to be a thought? If all it takes is possession of content, then, perhaps, we have supplied a recipe of sorts. But the product is pretty disappointing, a mere shadow of what we know (in ourselves and others) to be the fullest and richest expression of the mind. What I have described might, after all, be realized in a snail. What we want (I expect to hear) is something more, something exhibiting the complex dynamics, both inferential and explanatory, that our thoughts have. To have a thought about cows it isn't enough to have an internal, completely isolated, cow representation. To be a cow thought this representation must actually do what cow thoughts do. It must be involved in reasoning and inference about cows; it must (together with cow-directed desires) explain cow-related behavior. It must, together with cow-desires, rationalize cow-directed attitudes.

There is validity to this complaint. If we are going to make a thought we want the product to both look and behave like a thought. What we have so far devised may have some of the features of thought. At least it has a representational content or the sort we associate with thought. There is, however, nothing to suggest that our product will behave like a thought. Why, then, advertise the recipe as a recipe for thought? I have, after all, already conceded that there may be representations of this sort, mechanisms in the body having an indicator function, that are not *mental* representations at all. When the underlying functions are phylogenetic (e.g., in the processes controlling various reflexes), the representations are not thoughts. They have a content, yes, but they do not behave like thoughts. They do not, for instance, interact with desires and other beliefs to produce intelligent and purposeful actions. Why, then, suppose that when the functions are

ontogenetic, when they develop in learning, the results are any better qualified to be classified as mental? As genuine thought? An edge detector in the visual system might have the function of detecting edges and, for this reason, represent edges, but it is not—surely not for this reason alone—a thought about edges.

Since I have addressed this issue elsewhere,[17] I will merely sketch the answer. A system that acquires, in accordance with our recipe, and in its own lifetime, the power to represent the objects in its immediate environment will also, automatically, be an intelligent system, one capable of behaving in a rational way. To see why this is so, consider the process by means of which an indicator of F acquires the function of providing the information that F and, thereby, becomes a representation of F. In order to become the thought that F, this element must acquire the job of providing information about the F-ness of things. The only way it can acquire this function is by doing (i.e., causing) something—e.g., helping to bring about behavior A in condition F—that is beneficial to the organism when it wants or needs to do A. If the things this element causes are not useful or beneficial in some way, if they do not contribute to the satisfaction of the system's needs and desires, why should the element be selected to cause them? To acquire the function of indicating F, to become (thereby) a representation of F, therefore, a structure must play a part in the production of behavior that is rational from the point of view of the organism's well-being. An internal representation of F becomes a representation of F in a process in which what it causes is, in this sense, a *reasonable* response to F. According to this recipe for thought, nothing can become the thought that F without contributing to a rational response to F, a response that is appropriate given the system's needs and/or desires.

Something not only becomes the thought that F by assisting in the production of an intelligent response to F, it assists in the intelligent response to F precisely *because* it is signifies that condition F exists. That is, not only do thoughts that F conspire to produce intelligent reactions to F, they produce these reactions to F because they have an F-content. It is their content, the fact that they are F (not G, H, or K) indicators that explains why they are causing what they do. Had they been indicators of some other condition, a condition unrelated to a useful outcome, they would not have been selected for producing a response to F. This, it seems to me, vindicates, in one fell swoop, both the explanatory and rationalizing role of content. We do not need independent 'rationality constraints' in our theory of content. Rationality emerges as a by-product in the very process in which representations are created.

Our recipe, then, yields a product with the following features:

1. The product has a propositional content that represents the world in an aspectual way (as being F rather than G even when Fs are always G).
2. This content can be either true or false.
3. The product is a 'player' in the determination of system output (thus helping to explain system behavior).
4. The propositional content of this product is the property that explains the product's role in determining system output. The system not only does what it does because it has this product inside it, but it is the propositional content of this internal product that explains why the system behaves the way it does.
5. Though the system *can* behave stupidly, the normal role of this product (the role it will play when it is doing the job for which it was created in the conditions in which it was created) will be in the production of intelligent (need and desire satisfaction) behavior.

Our recipe gives us something that is beginning to both look *and behave* like thought.

NOTES

This article originally appeared with the title, "If You Can't Make One, You Don't Know How It Works" in *Midwest Studies in Philosophy* 19 (1994): pp. 468–82. I have made minor changes.

An early version was presented at the annual meeting of the Society for Philosophy and Psychology, Montreal, 1992. I used an enlarged form of it at the National Endowment for the Humanities Summer Institute on the Nature of Meaning, codirected by Jerry Fodor and Ernie LePore at Rutgers University in the summer of 1993. There were many people who gave me useful feedback and helpful suggestions. I am grateful to them.

1. I leave aside distracting complications having to do with the difference between magnetic and geographic poles.

2. Roderick M. Chisholm, *Perceiving: A Philosophical Study* (Ithaca, NY: Cornell University Press, 1957).
3. This is Searle's way of putting it in *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), p. 131, 156. As should be evident, I think Searle is wrong when he says (p. 161) that there are no aspectual shapes at the level of neurons. The sensory indicators in the brain are as much about the world (that we perceive) under an aspect as is the compass about the arctic under an aspect.
4. Despite even the *nomic* co-extensionality of 'F' and 'G' That is, a thought that x is F is different than a thought that x is G even if F-ness and G-ness are nomically related in such a way that nothing can be F without being G. This, too, is an aspect of intentionality. In Dretske 1981, p. 173, I called this the 2nd order of intentionality. Although compasses (indeed, all measuring instruments) exhibit the 1st order of intentionality (they indicate that x is F without necessarily indicating that x is G despite the co-extensionality of 'F' and 'G'), they do not exhibit the 2nd order of intentionality. If (in virtue of natural law) x must be G when it is F, then anything indicating that x is F will thereby indicate that it is G. If polar bears *cannot* live anywhere but the north pole, then compasses, in indicating the whereabouts of the north pole, will indicate the habitat of polar bears. Compasses cannot, while thoughts can, 'pry apart' nomically related properties.

   My discussion has so far passed over this important dimension of intentionality. It deserves discussion, but the complications are too great to cover in a brief article.
5. *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990).
6. In some cases, of course, an F will not even be *among* the causes of R since there are no Fs (unicorns, miracles, angels, etc.). This is a problem that, for lack of space, I skip over.
7. Louise Antony and Joseph Levine, "The Nomic and the Robust," in *Meaning in Mind: Fodor and His Critics* (Oxford: Blackwell, 1991), pp. 1–16.
8. Fodor 1990, p. 91, puts it a bit differently, but the point, I think, is the same: 'Solving the disjunction problem and making clear how a symbol's meaning could be so insensitive to variability in the causes of its tokenings are really two ways of describing the same undertaking.'
9. For the purpose of this essay I ignore skeptics about functions—those who think, for example, that the heart only has the function of pumping blood because this is an effect in which we have a special interest. See Searle's *The Rediscovery of Mind*, p. 238, and Dan Dennett's "Evolution, Error and Intentionality" in *The Intentional Stance* (Cambridge, MA: MIT Press, 1987).
10. This way of putting the point appeals to the view of natural functions advanced by Larry Wright in "Functions," *Philosophical Review* 82 (1973): pp. 139–168, and *Teleological Explanation* (Berkeley: University of California Press, 1976).
11. E.g., Ruth Millikan, *Language, Thought, and Other Biological Categories: New Foundations for Realism* (Cambridge, MA: MIT Press, 1984) and "Biosemantics," *Journal of Philosophy* 86, no. 6 (1989); David Papineau, *Reality and Representation* (New York: B. Blackwell, 1987), and "Representation and Explanation," *Philosophy of Science* 51, no. 4 (1984): pp. 550–72; Mohan Matthen, "Biological Functions and Perceptual Content," *Journal of Philosophy* 85, no. 1 (1988): pp. 5–27; and Peter Godfrey-Smith, "Misinformation," *Canadian Journal of Philosophy* 19, no. 4 (December 1989): pp. 533–50, and "Signal, Decision, Action," *Journal of Philosophy* 88, no. 12 (December 1991): pp. 709–22.
12. This may sound as though we are smuggling in the back door what we are not allowing in the front: a tainted ingredient, the idea of a *needful* system, a system that, given its needs, has a use for information. I think not. All that is here meant by a need (for system of type S) is some condition or result without which the system could (or would) not exist as a system of type S. Needs, in this minimal sense, are merely necessary conditions for existence. Even plants have needs in this sense. Plants cannot exist (as plants) without water and sunlight (they can, of course, exist as collections of elementary particles without water and sunlight).
13. This is the short and fast version of the story I tell in *Explaining Behavior* (Cambridge, MA: MIT Press, 1988).
14. Though we can encourage its development by artificial selection.
15. Fodor, *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990), p. 70.
16. I agree with Fodor (ibid., footnote 35, p. 135) that the only normative quality a naturalistic theory of meaning has to explain is the quality of being able to mean something that isn't so. If we can solve the problem of misrepresentation—or, equivalently, the disjunction problem—we will have all the normativity we want.
17. *Explaining Behavior* (Cambridge, MA: MIT Press, 1988).

Of Sensory Systems and the 'Aboutness' of Mental States

Kathleen Akins

Our thoughts, we believe, are 'about' things. When I look at a picture of the Eiffel Tower, I am thinking about an object, a particular structure that exists in Paris; when I remember the fragrance of gardenias, I recall a property that certain flowers have; and when I wonder how George Smith is getting along, I have in mind a particular person, a philosopher I met at Tufts University. Somehow or other, my thoughts are linked to properties and objects. The question, of course, is 'How'? How are mental events tied to the objects they represent? What is 'aboutness' and how does it come into being?

I want to discuss recent naturalistic theories of 'aboutness,' naturalistic theories of, roughly, how certain psychological or neural states come to represent, stand for, or be about properties and objects. The theories at issue are, among others, the theories of Patricia and Paul Churchland, Daniel C. Dennett, Fred Dretske, Jerry Fodor, David Papineau, Denis Stampe, and Kim Sterelny.[1] What I wish to show is that the naturalists' project, as commonly conceived, rests upon an intuitive and seemingly banal view of what the senses do—that the senses function to inform the brain of what is going on 'out there,' in the external world and in one's own body—and that this 'banal' view about sensory function is false or, more moderately, unlikely to be true in the strong guise that the naturalists' project requires. Hence I suspect that the naturalists' project, at least in its present form, is unlikely to work.

As one would imagine, much of this paper will be concerned with how I think the senses do function, in order to show exactly how the naturalists' project goes astray. The purpose of this paper, however, is not merely to cast empirical aspersions upon a philosophical project, to lob stones into the naturalists' camp, so to speak. Like most naturalists, I, too, believe that our thoughts/neural states are about the world—about the scent of gardenias, about the Eiffel Tower, about the tickles and itches in one's feet—and that this fact of 'aboutness' is fundamentally a *biological* fact about persons. Broadly construed, I share the naturalists' goal. It is the trodden path that has me worried. So the purpose of this paper is, first, to dissuade the average naturalist from the current route and, more importantly, to suggest a different path that a naturalistic theory of aboutness might take.

I. The Naturalist Camp and the Traditional View of the Senses

What, then, is the naturalists' project? Part of the problem with characterizing current naturalistic theories of representation is simply that, as a 'camp,' they comprise quite a diverse lot. That is, there is little agreement among the naturalists about what the project is—what exactly requires explanation—much less how the naturalistic theory should go. Take, for example, three of the more well-known naturalist programs, three among the many. First, there are those theorists, such as Fodor, who are committed realists, philosophers who take themselves to be explaining the content and directedness of our ordinary folk-psychological ascriptions—hence to be solving the traditional problem of the intentionality of mental states. On the opposite side of the clearing, one finds the eliminativists, with the Churchlands as their leading force. They deny that there are folk-psychological states, or more mutedly, they worry that the taxonomy of folk psychology will fail to survive the advances of the neurosciences. Eliminativists, then, are not concerned with the intentionality of folk-psychological states per se (for there likely are none) but with the problem of how *neural states* or *computational states* come to represent. Finally, a third distinct group, following Dennett, adopts 'the intentional stance.' On this view, the ascription of intentional states is a matter of holistic interpretation. When we ascribe psychological states to an individual, we are not describing her inner representational events. We are simply attempting to ascribe that set of intentional states that will have the greatest predictive power for the behavior of the individual concerned.

From *The Journal of Philosophy*, 93:7, 1996, pp. 337–72.

Qua elements of a predictive strategy, intentional states are 'real' but they are not the sorts of 'things' that stand in need of a naturalistic explanation. Still, Dennett holds, there are representations of some sort in the brain and it is the 'aboutness' of these states, whatever they might be, which will need a biological explanation. Hence the middle ground occupied by the intentional stance: a commitment to the taxonomy of folk psychology but a denial that folk-psychological states are the correct subject of a naturalistic theory. Clearly, then, the naturalists are divided by some deep philosophical differences: there is no consensus among them about what property, exactly, requires explanation—whether this is the property of intentionality, of psychological content more broadly construed, or of representational content, writ large. Nor is there agreement about what sorts of states have 'that' property—psychological states as delineated by ordinary ascriptions, psychological states as taxonomized by scientific psychology, neural representations, or computational states. Follow the sounds discord, emanating from the forest, and there you will find the naturalists' camp.

Their disagreements about the nature of mind and mental representation notwithstanding, all of the above naturalists share a common project. Each is trying to provide a naturalistic explanation of psychological representation and each is concerned (wholly or in part) with how, in the natural order of things, representational content comes into being—how a psychological state comes to be about a property or object. In an ecumenical spirit, then, I shall call 'the' property at issue *aboutness* and take the naturalists' project as one of explaining aboutness.

More specifically, most naturalists start with the view that aboutness is a relational property: for a representation to be about a property or object is for it to bear a specific kind of relation—call it the *aboutness relation*—to an object or property in the world. It is the nature of the aboutness relation, between a mental/computational/neural state and its object of representation, that the theory should explain.

Second, all of these theories are naturalistic in that they are motivated by the general conviction that our explanation of human psychology ought to be reconciled with our scientific understanding of the physical world. The goal is to explain this property of directedness as part and parcel of the physical world, as a natural phenomenon. In more concrete terms, this belief takes the form of a restriction on what will count as acceptable theoretic terms: the theory is limited to the terms of a natural science, a theoretic language devoid of semantic predicates.

Third, the naturalist usually begins by selecting a natural relation, one that will be used to explain the aboutness relation. Assume, for example, that a person is standing gazing at the Eiffel Tower. Those theorists who are attempting to explain the content of ordinary propositional attitudes might say of this case that the subject's perception is caused by the Eiffel Tower or carries information about it—they might posit an informational or causal relation between the subject's mental state and the Eiffel Tower. An eliminativist, on the other hand, who denies the categories of folk psychology writ large, might explain the event by positing a natural relation between a neural state and certain world events—perhaps it is the function of that type of neural state to indicate the Eiffel Tower or, more likely, to indicate towers in general. Naturalists agree, in other words, that, if a given mental/neural state is about x, there must be some natural relation (or relations) between that state of the individual and certain events in the world which ultimately explains this aboutness relation. What is contentious for the naturalist are two further issues: which natural relation ought to be chosen for this explanatory role and how, exactly, the posited natural relation (s) will be called upon to play that part. Will the natural relation be that of causation, of information, of indication, or will the natural relation be cashed out in terms of biological function? Is the intertheoretic relation, between the psychological realm and neurological/computational/biological realm, one of identity, of reduction, of supervenience, or perhaps something else? Here the views vary widely. What is generally agreed, however, is that the aboutness relation will be ultimately explained by a natural relation.

Fourth, virtually all naturalistic theorists agree on an important methodological point, namely, that a naturalistic theory should start with the static perceptual case. If we want to understand how our mental states come to be about the world, we should begin with those occasions on which our connection to the world is most clear cut—when, say, the subject sits, with his eyes open, staring at the world before him and has visual experiences of the forest beyond him. There is general agreement, in other words, that, if there are any simple cases of aboutness to be found, it is static perceptual events which

will furnish them. To put this another way, we all recognize that the relations of our ordinary mental states to their objects can be very complex. We have thoughts about things that are presently inaccessible to us by perception; we can ponder the nature of objects that are fictitious or abstract; we can think about things with which we are only vaguely acquainted, or which we are unable to identify, or about which we may have false beliefs. I can think about the Eiffel Tower even if I have never seen it; I can wonder how a certain George Smith is doing even though there are a number of philosophers named 'George Smith'; I can ponder the nature of causality or irrational numbers. These, we agree, are the difficult cases, the ones we ought best leave aside, at least until there is a story to tell about the simple case—when the object of perception is physically present to the subject and when he has some thoughts or representations that must surely be about that object.[2] It is the simple perceptual case that will, in the end, ground the aboutness of complex representational states.

The four broad assumptions, then, that naturalistic theories of aboutness usually espouse are these: that to give a theory of aboutness is to explain a relation between a representational state and some property or object; that this relation is to be explained in the terms of the natural sciences without recourse to semantic predicates; that this psychological relation will be ultimately explained by some relation from the natural sciences; and that the simple perceptual case is in some sense 'basic,' that it constitutes the most likely starting point for such a theory. This agreement, I think, is a surprising fact, given the diverse explanatory goals of naturalistic theories. If the goals of the authors are so heterogeneous, why do their theories have essentially the same broad theoretic form?

Enter the 'traditional view' of the senses, our intuitive understanding of sensory function. To present the traditional view in its strongest guise, think for a moment about the solipsistic plight of the brain. There floats the brain, the 'I,' encased and protected by the skull. It is the control center of the body, the mover of limbs, the planner of actions, the origin of all the body's thoughts and feelings. This central function notwithstanding, the brain resides in a kind of intracranial isolation. It has only mediated access, through its sensory and motor attachments, to its central concerns, to the body and the external world. But for a series of outgoing and incoming 'wires,' the brain is alone with its thoughts. It is because of the brain's solipsistic existence, then, that the senses have a clear role to play. They are the brain's window on the world. The senses show the brain, otherwise blind, how things stand, 'out there,' both in the external world and in its own distal body.

To flesh out this picture, consider one example of external perception, peripheral thermoreception, the system that reacts to surface skin temperature. Think about the sensations that we have of temperature—the bitter icy cold of a January wind, the pleasant coolness of a chilled drink in hand, the warmth of the spring sun, the singeing pain of a hot iron. From this internal evidence, one might guess that our thermal sensations are the products of skin receptors that are finely tuned to surface temperature. Like miniature biological thermometers, the receptors record the temperature of their immediate surround, its ups and downs. If, say, the skin temperature is eighty-nine degrees Fahrenheit, then the receptors must send a signal, some firing pattern, that 'means' eighty-nine degrees Fahrenheit; if the skin is cooler, say eight-three degrees, the receptors must give an appropriate response (perhaps they fire less rapidly). The receptors, we think, must react with a unique signal, one that correlates with a particular temperature state. Thus, the brain gains information about what it is like 'outside.'

Although we realize that human thermoreception probably does not function exactly as described above (for we all know about the various temperature illusions), the example captures a certain feature of the traditional view: sensory systems must be veridical in some sense of the word. If the senses are the brain's window on the world, then any system worth its salt (and functioning correctly) ought to provide an accurate account of just how things are: the brain must be able to tell, from the signals it receives, how things stand in the world.

First, it would not do to have a sensory system that is fickle or unreliable in its reports. No one would want a thermoreceptive system that suddenly started to send a signal, usually reserved for a skin temperature of eighty-nine degrees Fahrenheit, at arbitrary intervals; nor should the system be such that, when the skin *is* eighty-nine degrees Fahrenheit, the thermoreceptors fail to register that fact, forget to send the signal at all. What good, qua informant of the brain, would such a system be? In nonmetaphorical terms, this aspect of sensory veridicality is usually expressed as that of *constant correlation*: if a signal is to be informative ('tell the truth'), it

must be produced when and only when a particular stimulus (or stimulus set) is present.

Second, if the brain is to have an accurate understanding of how the world lies, then constant correlation is not enough. The relevant *structure* of the external events or properties must also be preserved by the sensory signals: the representational relations among the sensory signals must mirror the relevant relations in the sensed domain. In the case of thermoreception, the brain must be able to discern a one-dimensional relation between temperatures—whether this temperature is greater or less than some other one. So a thermoreceptive system must be, well, thermometer-like: the relations among the temperature states in the world ought to be at least roughly discernible from (if not strictly isomorphic to) the relations among the individual thermoreceptive signals. Constant correlation alone is not very helpful, especially if the question before the brain is whether or not to pull a hand away from the warming fire (that is, is the hand getting too warm now?).[3]

Third, a veridical sensory system is also a *servile* system (or if you prefer, a system that acts as the brain's loyal retainer). By 'servile,' I do not mean a system that reacts in only strictly prescribed (law-like) ways to the world's sensory impingements. Much of sensory processing, we now believe, involves active or 'top down' processing: the brain uses stored information, default assumptions, or hypothesis generation and testing in order to solve the computational problem at hand ('What is that object'?) or to increase its efficiency. Quite obviously, this kind of activity does not challenge our intuition of veridicality: such sensory systems are still trying to represent the world, in the face of limited information, as accurately as possible. What would not jibe with our notions of a proper (veridical) sensory system is one which actively embroidered upon the nature of the impinging sensory stimuli or which simply made things up. What we expect from sensory systems, in other words, is that they are the brain's 'ontological drones.' In the service of the brain, they toil tirelessly to report the 'what, when, and where' of the world's events. They do not interject their own opinions into their reports; they do not slyly skew the information to reflect their own interests or prejudices. Their job is to state the facts. To say that a veridical sensory system must be 'servile,' then, is to say that it represents the world as accurately as possible, without embroidery or fiction, given the information available. (I take servility to be a slightly different notion than that of reliability as given above, though perhaps it is merely a subtype. By analogy, the problem with the town gossip is not that, for any event in town, she might fail to have something to say about it. What we doubt is the *accuracy* of what she says. One worries that it will incorporate her own likes and dislikes, prejudices, and interests.)

On the traditional picture, then, the senses, using a system of signals that captures the structure of a domain of external properties, tell the brain, without exaggeration or omission, 'what is where.' It is this view of the senses which dovetails with the philosophical problem at hand. The naturalists' project, stated in its most general form, is to explain the psychological relation of aboutness in some way that fits neatly with our scientific picture of the world. On the traditional view of the senses, there is also a relation of aboutness, a relation between any state of a veridical sensory system and its object. Indeed, fulfilling this relation is the raison d'être of the senses—that without which the brain would not know how things stood, either in its own body or in the external world beyond it. So what the traditional view of the senses provides, from the natural sciences, is exactly the sort of relation a theory of aboutness requires—hence an obvious starting place for any naturalistic theory of aboutness. Not all intentional states, the naturalist freely admits, are sensory states; nor need the subject have ever had sensory contact with a given object of thought; nor indeed are the intentional objects of such states necessarily physical objects or properties, the sort of entities with which the subject could have sensory contact. Still, the path to choose is clear, the naturalist will contend. Indeed, at first glance, it is the only path visible at all—the only path out from the solipsistic existence of the brain.

II. Narcissistic Sensory Systems

The problem with the traditional theory of sensory processing, I think, is that it is not universally or even generally true. In some cases, sensory mechanisms do behave as one intuitively expects, in accordance with the traditional view. More often than not, however, sensory systems fail to be 'veridical' in the sense given above. Indeed, if one had to pick a single predicate to describe all sensory systems,

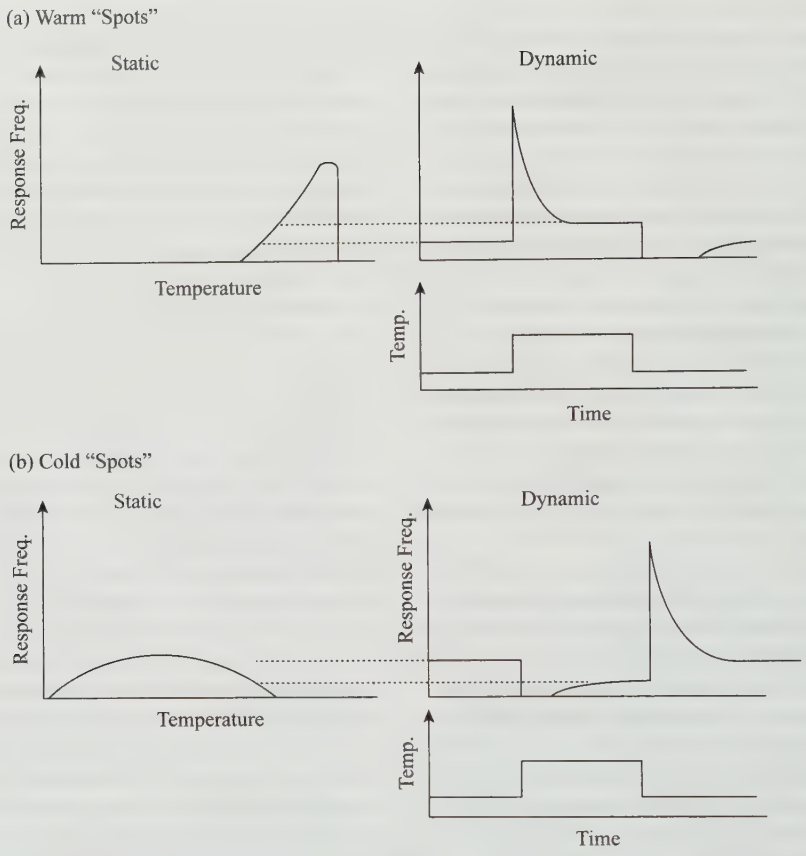
it is that each and every sensory system is, well, 'narcissistic.'

As a first pass at explaining this metaphor, think of the usual case, the human narcissist, a person whose worldview is informed by exactly one question: 'But how does this all relate to ME?' In a classic story, a narcissist goes to his therapist for his regular appointment. At the door he is met by another therapist in the same practice: there will not be a session that day, she informs him. The narcissist's therapist has been in a boating accident—she is alive, but in critical condition in the hospital. It is gently explained that there is a good chance that the therapist will survive and recover, lead a reasonably normal life, perhaps even return to her practice. At this, the narcissist looks stricken, and wails: 'But why do these things always have to happen to ME?'

Needless to say, a narcissist's world picture, being informed by but one question, is strangely askew. In this regard, it is important to realize that the problem with the narcissistic worldview is not just that his range of interests is idiosyncratic, that self-interest directs his attention toward a limited portion of the world, a small part of what other people see. Rather, by asking the narcissistic question, the *form* of the answer is compromised: it always has a self-entered glow. In this case, it is obvious that the causal factors leading up to the accident are entirely independent of the narcissist's existence. But it does not seem this way to him. The narcissist cannot see himself and his relation to the world in an objective light, as one life among others. Hence his understanding of the emotions and actions of other people, and of events in the world in general, will necessarily incorporate his own particular interests. For the most part, it is not possible for the narcissist to stand back and remove himself from the picture—and that, of course, is exactly the property which gives the narcissist away. I do not mean to imply by this that the narcissist never gets things right, never sees the world for how it is. Sometimes the question 'How does all this relate to me?' is just the right question. 'What does that person want from me now?' is the appropriate thing to wonder in a dark alley. Hence it yields a legitimate—'veridical'—answer. It is nonetheless true that the narcissist's question shapes his worldview at all times, and this is so regardless of whether the question is 'appropriate' or not.

Turn again to a simple sensory system, that of thermoreception. As I said above, our first guess about thermoreceptors is that they act like small thermometers, signaling the brain about the location and the temperature state of the skin. In fact, however, the nature of thermoreception is quite different (the following draws largely upon from the work of H. Hensel[4]). First, the apparently continuous temperature gradient that we feel is not the result of the continuous response of a single thermomechanism. Instead, our sensations are the result of the action of four different types of receptors: two thermoreceptors, 'warm spots' and 'cold spots,' and two pain receptors (nociceptors) that fire only in the extreme conditions of very high or very low temperature. At very high and very low temperatures, we feel only pain, sensations that are qualitatively indistinguishable from one another. In the middle zone, we rely upon one or the other kind of thermoreceptors for our sensations of warmth and cold. (The temperature at which warm spots begin to respond is roughly the same temperature at which cold spots leave off.) Second, there are far more cold receptors than warmth receptors, the exact ratio differing from location to location. So, for example, on the face, the nose has a ratio of 8:1 cold to warm spots, the cheeks and the chin are somewhat more sensitive to warmth with a ratio of 4:1, and the lips (entirely counterintuitively) are sensitive to only cold, with almost no warm spots whatsoever. One result of this variability is that different parts of the body are more sensitive to heat or cold than others. Because conscious sensation is the result of cumulative neural response, the more receptors there are, the more cumulative neural activity. This is a fact that will strike you as immediately plausible if you imagine wading into a cold lake. As a matter of fact, some steps *are* harder to take than others. Another result of this variability is that the 'temperature of neutrality'—the 'comfort zone' as thermostat makers call it—differs from one area of the body to another. Whether a temperature feels comfortable is as much a function of location on the body as it is of temperature.

What do the individual receptors do? Each kind of receptor, warm spots and cold spots, have both a static and a dynamic function (figure 1). The receptor's response to a constant temperature is its static function, represented by a curve that plots the rate at which the neuron fires against the stimulus temperature. For both the warm spots and cold spots this response is nonlinear; the static functions for the two receptors also differ, one from the other. The



Figures 1(a) and 1(b). The static and dynamic functions of the warm and cold receptors. To illustrate dynamic function, both receptors were subjected to a sudden temperature increase and then a sudden temperature decrease. The warm receptor shows a dynamic response to temperature increase alone; the cold receptor shows a dynamic response to temperature decrease alone. (Adapted from Hensel, *Thermal Sensations and Thermoreceptors in Man.*)

warm receptor responds over a narrow range of temperatures with a steep rise in firing rate at the high end; then the firing abruptly halts at a maximum temperature. The cold receptor has a less intuitive response pattern. It has a wider window of response, a gentle curve with a maximum response at the midpoint, tapering off thereafter. The static functions of neither the warm spots nor the cold spots are thermometer-like, with a certain set increase in firing rate per degree of temperature change.

Both kinds of thermoreceptors also have a dynamic function, a response to temperature change. When the temperature of a warm spot is increased, a burst of activity occurs; then the firing rate gradually slows and settles into a new higher base rate, a rate determined by the static function. For example, after shoveling a snowy walk, putting your hands under only tepid water initially feels very warm. The dense liquid causes sudden energy transfer, a sudden increase in temperature, and the dynamic function's burst of neural activity. When the temperature of your hands stabilizes, the neural activity also decreases and the water feels as it normally would, cool. What is important here is that the size of the initial activity burst is *variable*: it depends upon the starting temperature. So, for example, if you transfer your hand from tepid to warm water—say, a forty degree-Fahrenheit change in temperature—there will be a sudden burst of activity in the warm receptors. But if you then transfer your hand from warm to hot water, again a difference of forty degrees Fahrenheit, the dynamic burst will be much greater. The higher the starting temperature, the larger the initial burst of the warm receptor. Thus, the transition from warm to hot water is felt more keenly than the change from tepid to warm water. Note also

that when a warm spot is cooled there is no neural burst at all, only a gradual decrease in firing rate until the new lower base rate (determined by the static function) is reached. These same principles apply to cold spots as well, except in reverse. A *cold* stimulus evokes a dramatic over-compensatory response; a *warm* stimulus causes a gradual decrease in activity; and the *lower* the initial skin temperature, the greater the initial burst of activity in response to a cold stimulus. (These opposing characteristics are, in fact, the defining functional qualities of warm and cold receptors.)

As one can see, the response properties of the thermoreceptors are complex. There is no single thermometer-like receptor or even two 'abutting' thermometer-like receptors; the warm spots and cold spots exhibit two different nonlinear (static) functions; the transducers exhibit adaptation and behave in a context-dependent fashion (the dynamic response depends upon the current base-line skin temperature). What is more, our temperature sensations do not result directly from the functional properties of individual receptors but are determined by the characteristics of the neural population as a whole. Where there are more cold receptors, the area seems colder; if a location has few warm spots, heat is hardly noticed at all.

All of this seems somewhat strange on the traditional view of sensory processing, of thermoreception as a system that disinterestedly records temperature facts. Just how inept could this system be? Viewed as narcissistic, however, the system makes perfect sense. What the organism is worried about, in the best of narcissistic traditions, is its own comfort. The system is not asking, 'What is it like out there'?—a question about the objective temperature states of the body's skin. Rather, it is *doing* something—informing the brain about the presence of any relevant thermal events. Relevant, of course, to itself.

In this light, reconsider the old illusion created by placing one hand in cold water, the other in hot, and then, and after a few minutes, placing both hands simultaneously in some tepid water. Stupid sensors. They tell you that the tepid water is two different temperatures. But the sensors seem less dull-witted if you think of them as telling you how a stimulus is affecting your skin—that one hand is suddenly cooling while the other is rapidly warming. Skin damage can occur from extremes of temperature (being burnt or frozen) but also from rapid temperature changes alone, even if the changes occur within the temperature range for healthy skin. So the information provided by the dynamic function—whether there has been a change in temperature, in which direction and how rapidly that change is occurring—is crucial to the organism for avoiding skin damage. It allows you to pull away your hand before it is burnt, to remove your hand when the change in temperature will be fatal to the skin.

From this perspective, one can also see why it makes good sense to have a dynamic function that is context dependent—why the burst rate depends upon both the amount of change and the starting temperature. If the skin temperature is already high, any increase is likely to cross the upper limits of skin tolerance. So the extreme bursts of the dynamic function, produced by stimuli at the upper end of warm-spot function, warn that tissue damage is about to occur. (If the temperature sensitive nociceptors fire, then damage has already been done.) In other words, the system produces strong signals whenever potentially dangerous temperature changes occur. From this perspective, the utility of the dynamic function is clear.

As another example, think again about a cold mountain lake—this time, about putting your head into a cold mountain lake. Most of us have had the experience of going for a hike and then, after a strenuous day, coming upon an inviting-looking lake. First you put your hands in the water to test the temperature; then, after the initial shock of is over (after the dynamic burst of the cold spots has subsided), you decide to risk a swim. Note, however, that making your brain very cold is not a good idea. Heat loss through the scalp is one of the many ways that human life comes to an end. So, from an evolutionary perspective, it is not surprising that the scalp has an abundance of cold receptors—that when you dive in, the water feels colder to your head than it did to your hands earlier. Given how dangerous heat loss can be, it makes sense to receive a sharp warning signal whenever the scalp is cooled, before you lose the sense to react rationally. More generally, if one looks at the complexity of the human body and at the large range of behaviors we are capable of performing, it is clear that the importance of thermal stimuli can vary, depending upon where the stimuli occur. You need to keep the top of your head warm; you need to make sure that your nose is not frozen in weather where your cheeks will still be fine; you need to be able to sense both cold and warm objects with your hands. Hence the evolutionary 'solution' of a variable

ratio of cold to warm spots. Through a difference in the number of cold and warm spots, the cumulative signals are tailor-made to reflect the relevance of temperature stimuli for a particular place on the body.

Looking back at thermoreceptive function, then, one realizes that this system is not merely inept, a defective indicator of surface temperature. Rather, the system as a whole constitutes one solution to man's various thermal needs—that he be warned when thermal damage is occurring or before it is likely to occur, when temperature changes are likely to have specific consequences, and so on. These are the thermal problems which have been 'posed' over time to the evolving organism and which have resulted in the idiosyncratic functions described above. In other words, this system, qua system, reflects a constellation of interacting facts about one species—the environmental conditions encountered by Homo Sapiens (for example, the thermal properties of the environment), the developed physiology of the species (for example, the mortality conditions of the skin, the details of internal and external thermoregulation), and the range of behaviors of which the species is capable.[5]

Let me now make explicit how the above characterization of thermoreception fails to fit the traditional view. Recall from the ways in which, according to the traditional view, sensory systems are veridical: each signal must correlate with some property (or range of properties) in the world; the structure of the relevant relations among the external properties must be preserved in a systematic encoding of those relations; and a sensory system must be servile in that it must be seen to be reconstructing, without fiction or embellishment, the properties, objects, and events of the world external to the brain. In this case, the property of interest is, obviously enough, skin temperature at a specific part of the body. So the question that one would expect the thermoreceptive system to answer is: 'What is my skin temperature at *x*'? But this is not what it does.

First, particular thermal sensations do not necessarily correlate with any particular temperature or any particular temperature change. Because thermal sensations are a function of the firing rates of a neural population and because the absolute number and ratio of the two different receptors differ from one part of the body to another, exactly the same skin temperature can give rise to a variety of sensations. This is one reason. For another, the elliptical static response properties of cold receptors ensure ambiguity in its signals. As the skin becomes colder, a cold receptor fires more and more rapidly until the receptor's maximal response level is reached; then, as the temperature continues to drop, the firing rate starts to decrease. So a single cold receptor will fire in exactly the same way for two different stimuli, a temperature at the low end of its response range and a temperature at the high end. Necessarily, its signals (alone) are ambiguous. Nor do thermal sensations reflect temperature change. Here, it is the context dependency of the dynamic function of both warm spots and cold spots which presents the problem. The felt change in temperature for a specific temperature change will depend upon the starting temperature of the skin. If the temperature of a warm spot is increased at the bottom of its response range, the dynamic burst will be small; if it is warmed at the top of its response range, the burst will be very large. So neither absolute temperature nor temperature change is recorded by thermal sensations.

Second, for almost the same reasons, thermal sensations as a whole do not reflect the structure of thermal stimuli, whether stimulus T1 is greater than, less than, or equal to stimulus T2. The water, as we wade into it, initially lessens the skin temperature of each body part about equally; but it certainly does not feel that way. Some parts feel much colder. Similarly, cold receptors, with their elliptical static response function, represent disparate pairs of temperature stimuli as being equal—very low ones and relatively high (tepid) ones. Even warm receptors, with their nonlinear static responses (which abruptly stop at a certain maximal rate), do not quite get the picture right. Lower temperatures do elicit lower firing rates but the differences in temperature between the warmer and hotter stimuli are not uniformly recorded. Linear temperature increases are encoded as nonlinear changes.

Third, the thermoreceptive system embroiders its account of the temperature states of the world. Unlike many other sensory systems, it is probably unfair to say that this particular system actually manufactures fictions in the course of its ordinary function, but it does appear prone to chronic exaggeration. At the lower and upper limits of response for the cold spots and warms spots, respectively, a small temperature change elicits a hysterical response. Moreover, given the differing ratios of cold spots to warm spots, no body part simply reports its surface temperature. Each body part exaggerates its own state in accordance with of its own interests and

sensitivities. So the thermoreceptive system is hardly the ontological drone that the traditional view had imagined, the tireless reporter of the 'what, where, and when' of surface temperature.

It is time to step away from the details of human thermoreception in order to make some more general points about sensory function. In imagining the function of any sensory system, we slide very easily from the question 'To what are the receptors responding'?—a question about the nature of the proximal stimulus that will evoke a receptor response (for example, mechanical deformation or light in a certain range of wavelengths)—to the question 'What do the signals of the system detect'? In the grip of the traditional view, we treat these as one and the same question. But despite the entirely rule-governed nature of sensory systems, the question 'What is the system detecting?' may not be apt. Rather, the appropriate question, for any sensory system, is 'What is the system doing?' and by this we mean doing for the organism. This point requires some explanation.

Each and every sensory system, no matter how sophisticated or simple, is tied to a set (sometimes a very large set) of behavioral tasks. No matter what else the senses do, in the end, they must inform movement or action. In the case of behaviorally simple creatures, there may be a very short route from the response of the sensory receptors to the output of the motor system. In the slug, no doubt, a dynamic burst in a population of receptors sensitive to sodium chloride will trigger an immediate aversive reaction, movement away from the stimulus. Here, the utility of matching the information recorded by the senses to the needs of the motor system is clear. There is no need to represent the world 'the way it is,' for simple lives (behavioral repertoires) require limited information. Indeed, evolution will favor sensory solutions that package the information in efficient and quickly accessible formats, in ways that match the particular physical form of the motor system, its motor tasks, and hence informational requirements. For every evolved system, there will be a symbiotic relationship between the information gathering of the sensory system and the informational needs of the motor system—and the elegant solutions that evolution eventually selects need not involve any straightforward (to our eyes) 'veridical' encoding of sensory information.

In our own case, of course, we think before we act. We are not simple stimulus-response organisms, with our behaviors triggered by simple sensory signals. Moreover, our perceptual thoughts are genuinely intentional, tied as they are to the objects and properties of the external world. I sit and stare at my empty coffee cup; I wonder whether I would like another cup, or at least, want one enough to get up and make one; I ask my husband whether he would like a cup of coffee. (And so on—philosophical drama at its finest.) These two facts, that our responses to the world are not stimulus driven and that our intervening thoughts are about stable objects and their properties, may lure us into thinking that somehow human sensory systems (and those of other behaviorally complex creatures) are of a different, 'veridical' sort. Our senses could not be narcissistic, or at least, not for the most part. (Let us dismiss thermoreception as some crude and embarrassing vestige of our primitive past.)

Neither of these two facts, however, obviates the need to direct motor behavior in a timely and efficient fashion by means of sensory information. Ruminate about coffee all you like, but should a full cup appear, you will have to reach out your hand and grasp the cup. From your perspective, of course, the point of view of the conscious subject, it may seem to you as if you look out at the world, contemplate the cup of coffee on the table, and then decide to pick it up—as if your unitary perception of the coffee cup qua cup itself unmediatedly directs your hand. After all, if the result of all those visual processes is a perception of a cup, surely that (unified) perception then informs your actions. No matter. What directs your movements, from the moment you begin to lean forward in anticipation of reaching with your arm to the moment the coffee cup makes contact with your lips, is a host of information from numerous different sources. There will be *visual information* about the egocentric position of the cup relative to your body, about its position relative to your reaching hand, about the shape of the handle relative to your grasp, about the cup's rotation (shape) relative to a horizontal plane as you pick it up (do not spill it now), and about the cup's speed of movement; there will be *proprioceptive information* about the position of your upper body as you lean forward, about the angles of your arm joints as you stretch toward the cup, about the weight of the cup relative to the firmness of your grip, about the fantastically minute adjustments of your fingers, hand, and arm muscles as you balance the cup of liquid in an upright position, about the position of the cup and your hand relative to your lips (after

all, you do not have to stare cross-eyed to get the cup there); there will be *tactile information* about the pressure of the cup in your hand, the pressure of the cup on your lips, the shape of the cup in your hand. No doubt this is but a small part of what is actually involved in the 'simple' activity of picking up a cup of coffee. Even our simplest actions, then, involve numerous sources and types of information (here, visual, proprioceptive, and haptic information) and, within a single system such as vision, specialized information (about shape, position using a variety of reference frames, rotation, movement, and so on) which requires diverse representational schemes. Thus, if one looks at the neuroanatomy of mammalian vision, most of the physically distinct sites, each associated with a specific informational problem, have connections with one or more motor site(s) and all visual areas have connections with subcortical sites which themselves have lines to motor areas. Diverse and complex information subserves simple motor movements. We, as perceivers of cups and saucers, then, tend to mistake our conscious perspective, about coffee and cups, for insight into how things work, how we actually manage to raise the cup. But this is not how it is. Although, from the first-person perspective, what you see is a cup sitting on the table, what actually guides your movement is a plethora of sensory signals. The upshot is that for us, as much as for other creatures, the symbiosis between sensory information and motor needs is equally strong. If the sensory information is to guide the requisite motor movements, it must be usable on-line, by numerous feedback systems, control loops, and cognitive 'interrupts.' All the sensory information must be encoded in a motor-friendly way. Even as intentional, conscious perceivers, we are equally in need of narcissistic sensory strategies, indeed, perhaps more so given the complexity of our behavior and bodies.[6]

When we examine a sensory system, then, we are looking at an evolved solution to a specific informational problem. Perhaps it is a system that is able to increase or decrease the sensitivity of its receptors in response to present needs (without recording the change in sensitivity), that applies filters for the creation of specific features, that adjusts an outgoing signal in order to ensure certain characteristics in the signal return—or that encodes the facts about skin temperature in some way designed to aid creature comfort. The relevant question to ask of such a system is 'What is it doing'? This is a question to which the answer might be 'it is measuring, with variable discrimination, the animal's tilt away from the vertical in order to maintain its upright posture,' 'it is providing a visual signal that can be easily processed for movement information,' 'it is monitoring the stretch of the flexor muscle in order to adjust the length of the tensor by an equal amount,' or 'it is indicating an edible insect'—answers that may or may not make reference to processes of veridical perception.

III. Defending the Traditional View: Three Objections and Replies

In the face of the above kind of neurophysiological example (for human thermoreception is merely one among many, chosen for its familiarity), there are a number of responses one might give in order to vindicate the traditional view. I shall give three of the most common replies. The first response is, in effect (but not in intention), an a priori defense of the traditional view; the second response is a straightforward denial of the thesis that sensory systems are, at bottom, narcissistic; the third response, the most promising of the three, grants that sensory systems are narcissistic—but it denies any conflict with the traditional view.

A. The a priori defense. Given the above example of thermoreception, one might take away quite a different lesson, not a lesson about sensory processing per se, but a metaphysical view about the nature of *properties*. Dennett, for example, in talking about our conscious perception of secondary properties, takes the clearly self-interested nature of sensory systems to show that their normal function can *define* properties in the world. It does not matter that our sensations of, say, redness fail to correlate with any easily delineable set of surface reflectance properties, a single property that is sanctioned by the scientific image. There is redness in the world because we have a disposition to respond in a 'discriminating' fashion to 'red' stimuli.[7] That is, if a sensory system responds in one and the same way to a rather diverse set of properties and conditions, then that disorderly set simply *is* a property of the world. I am not sure that Dennett himself would wish to frame the view in this general form, but let us assume this strong view for the sake of argument.

Setting aside the merits of this view qua metaphysical thesis about properties, note that as an answer to the question 'Are sensory systems veridical?' it is empty. If properties are defined by the ordinary causes of sensory signals, then, by definition, all sensory systems are veridical: the system 'captures' the structure of the domain of external properties just because it *defines* that domain and, of course, any signal reliably records, without exaggeration or omission, whatever stimuli ordinarily cause it. It is only against a firm prior ontology of properties, one that is specified and justified independently of actual sensory response, that questions about veridicality have any bite at all. Thus, on this reply, what seemed like a proto typically empirical question—'Are the senses veridical?'—is recast as a question of definition, the answer to which has the status of an analytic truth. This alone, of course, is not a reason to dismiss the response—perhaps the traditional view ought to be granted the status of analytic truth—but this seems to me unlikely. (Below, in answering the third objection, a more decisive objection to this response will emerge.)

B. The appeal to signal information. Whether or not a sensory state appears to correlate with a particular external property, this response begins, sensory states nonetheless *carry information* about their causes. Take the case of thermal sensations. Although any given thermal sensation might be caused by a large variety of temperature states, a given thermal sensation still carries information about its cause. It just does not wear this information on its sleeve. After all, thermoreceptors act in orderly ways—the response functions of both warm spots and cold spots are shown in figure 1 and there is some function that sums the responses of individual receptors. So starting at the beginning of the process, with the initial state of the thermoreceptive system (the current static response rate of the receptors), plus the neural population characteristics (the ratio of warm spots to cold spots and their numbers) and stimulus temperature of the skin, one could predict the sensations of the subject. Conversely, starting with the resultant thermal sensation (or population response), plus the initial state of the system and the neural population characteristics, one could compute the value of the stimulus temperature. In other words, it is possible, using the appropriate calculations, to deduce skin temperature from the thermal sensation under standard conditions—and this is just what it means to say that a signal (thermal sensation) carries information about a source (skin temperature). Hence, in this sense, an informational sense, thermal sensations do reliably indicate temperature states.

There is certainly some truth to this response. Given the causal regularities that govern sensory systems (for there is nothing supernatural about them), the sensations or signals they give rise to will often carry information about their causes. That is, whenever there is a computable function that describes the input-output relation, the response of a sensory system will carry information about its causes. In the case of thermal sensations, sometimes our sensations carry information about skin temperature and temperature change, sometimes not. This is because the static response of the cold receptors does not define a computable function, hence there is no algorithm that could determine, given the population response of the cold-spots (a sensation of cold), the stimulus temperature.[8] Indeed, a bell-shaped response curve for sensory receptors is more the norm than the exception, but let us put aside this fact for the moment. The question that concerns us here is whether, given that information about the stimulus is often carried in the sensory signal, this will be of any practical use in constructing a theory of aboutness.

The problem here is that it makes little sense to identify the contents of sensory states with whatever information they carry. Return to the example of wading into the lake. There you stand in water up to your thighs, debating the relative advantages and disadvantages of total submersion. (Would not a bit of splashing and swishing be just as effective, you wonder?) Eventually, you wade in, gasping and sputtering at various intervals. Now, on the informational view, all this silliness is for naught. The dynamic bursts of the various cold-spot populations, no matter where they may lie, all signal exactly the same thing—a single temperature for the water or, if you like, a single temperature change. This is the information carried objectively by each population of cold spots—information that you already had when you waded in up to your thighs. Of course, *you* may behave a little oddly, as you immerse yourself, but there is nothing odd about the behavior of your thermoreceptive system. It reliably indicates a single change in skin temperature, no matter what parts are presently getting wet.

This, of course, is a parody of the information response, but it makes a serious point. Information that is carried by, but not encoded in, a signal is information that is available

only *in theory*. To say that the information is *present* is to say only that there exists a computable function that, if used, would yield the correct result—in this case, the actual temperature change. It is present, as it were, from the point of view of the universe. But no creature has ever acted upon information that is available only in principle. Thus, to posit a theory of aboutness based upon signal information alone is to a posit a theory that, by hypothesis, makes no connection with an organism's action, behavior, or thought. I take it that this is not what the objector had in mind.

Rather, when an objector points out that even a narcissistic response contains information about its causes, the hope is that somewhere down the line, this information is *extracted*. Despite our conscious narcissistic sensations or the narcissistic responses of sensory systems, somewhere higher up in each sensory system the appropriate calculations are made and the true state of the world is inferred. (Thus, for example, in case of human color vision, the ambiguity inherent in the bell-shaped response curve of each cone—ambiguity between intensity and frequency information—is partially resolved by the overlapping range of responses of the three cone types. Through comparison, ambiguity is resolved.) How else could a sensory system be of use? This suggestion, however, amounts to little more than an expression of one's faith in the traditional view. Empirically, there is little reason to think that all sensory systems carry within them the means to 'decode' their own responses.

Once again, in the case of thermoreception, there is no anatomical or physiological evidence to suggest that the thermoreceptive system has the additional information needed to compute stimulus temperature—that it has knowledge of the number of receptors at each location, or the nature of the static and dynamic functions of the receptors, or that it keeps track of the system's initial state. Nor is there any further psychological or behavioral evidence that our thermoreceptive systems have this information or make these calculations. Given a thermal sensation alone, we have no capacity consciously to 'see through' the narcissistic thermal sensation to the objective skin temperature (or temperature change) nor, behaviorally, do we act as if (unbeknownst to us) our brains make these kinds of calculations. This information is simply not available to us through sensory means. (Of course, there are many other ways that one might use thermal sensations to infer temperature. After reading Dr. Spock, you will know what temperature of milk correlates with a certain warm feeling on your wrist. Knowing this, the sensation can be used as an indicator of a particular milk temperature. But to make this inference requires exactly those cognitive capacities which the naturalistic theory hopes to explain—intentional thoughts about Dr. Spock, baby books, wrists, sensations qua sensations, and so on. The thermoreceptive system alone does not provide this information.) The point, here, is a general one: when a sensory system uses a narcissistic strategy to encode information, there need not be any counteracting system that has the task of decoding the output state. If the very point of a narcissistic encoding is to match the incoming information to an organism's behavioral needs, then for most part, there is no reason to re-encode the sensory information into a veridical format.

C. The appeal to detector cells and biologically salient properties. Of the three objections, the third is the most promising as well as the most obvious: Granted that sensory systems utilize narcissistic encodings, why not say that the function of a sensory system is to detect *narcissistic properties*, properties that are defined relative to an organism's interests? The shocking cold of the lake that you feel on your scalp, for example, is a property, one that *that* sensation (population response) detects, namely, the property of being-too-cold-for-my-head. The same will hold for the properties detected by similarly narcissistic mechanisms. To put this another way, when the question 'What temperature states do thermal sensations indicate'? was asked, the answer was restricted to the temperature readings of some scientific scale, degrees Fahrenheit or Celsius or Kelvin. On this division of the world, the answer was 'no': one thermal sensation does not indicate any single skin temperature (or temperature change). But if we look to the neuroethological or neurophysiological literature on sensory function, we find descriptions that make use of a variety of biologically salient properties, properties that are not described by the predicates of physics and chemistry. We find, that is, descriptions of 'detectors' in simple organisms, some of which respond selectively to 'legitimate' properties, such as magnetic north, being a complex sugar or a certain amino acid, or having a certain wavelength, others of which respond to 'messy' properties, such as vertical symmetry, small flying insects, movement in the left of the visual field, or being a poisonous substance.

Moreover, referring to these messy properties is essential to characterizing the function of these detectors; they are an ineliminable part of the neurobiological description. Thus, if we recognize 'narcissistic' properties, along with 'legitimate' and 'messy' properties, as biologically salient, we can recast sensory systems in a way that conforms to the traditional view—as detectors or reliable indictors of external properties.

What makes this a particularly good objection, of course, is that much of it is right. I take it as incontrovertible that there are neurons that act as detectors, and that these detectors are tuned to properties such as predator and prey. Biologically salient properties are exactly those properties which neuroethology and neurophysiology appeal to in describing sensory function and those without which sensory function qua function could not be characterized. Moreover, prima facie, there is no reason to disbar narcissistic properties from the club, if neurobiological description demands them. Admitting all of this, however, does not give the critic what he needs.

Notice, first, just how strong the suggestion is qua neurobiological thesis. The claim is that each and every sensory system functions to detect properties, be they narcissistic properties (defined relative to organism's needs), biologically salient 'messy' properties (for example, the property of vertical symmetry), or 'legitimate' properties (those recognized by the other physical sciences, say, the property of containing NaCl). Call this the *detection thesis*. This is an extremely strong universal claim about sensory function and hence about the form of all our future biological explanations of the senses. Moreover, it is a universal claim made in advance of the lion's share of empirical research on, and biological theorizing about, a vast topic—a claim, the only present empirical basis for which could be a few notorious examples from the comparative neurobiological literature (for example, the famous fly detectors of the frog). It is a testimony to the strong intuitive pull of the traditional view, I think, that the prematurity and all-encompassing nature of the detection thesis seems to have escaped most of its backers.

Note that, even if one were to take all of the present literature on sensory processing and think up narcissistic or messy properties for those systems to detect, this would not provide any evidence for the universal claim. There are, after all, many devices, natural and human-made, for which one could claim a rough function from input to output states, for example, for vacuum cleaners and stereo amplifiers, as well as lungs, livers, and intestines. With a bit of imagination and a good sense of humor, one could think up some 'messy' or 'narcissistic' properties for these devices to detect. What the objector is claiming, however, is not merely that we could conjure up properties for each system to detect. He is making the stronger claim that neurobiologists will usefully describe the systems as such—that characterizing sensory systems as detecting properties will *always* provide us with some further insight into how or why the mechanisms function as they do. In other words, the detection thesis is committed to the view that our best biological explanations will characterize sensory systems as detecting. As I said above, this is a very broad empirical claim, especially when one realizes the infant state of neurobiological research on the senses. On what grounds might someone believe it to be true?

Above, in giving the example of the human thermoreceptive system, I argued that the neurobiological research does not support the traditional view nor, for that reason, the detection thesis—that we must ask the broader question 'What is the system doing'? not merely 'What is the system detecting'? In the literature of that nascent science, one finds, along with references to the now-famous detector cells, a variety of other descriptions. There, the neurobiologist begins his research by trying to pinpoint the informational problem(s) that the system must solve (a good alternative to poking blindly about) and then looks for mechanisms that solve them. Here, because the problems and answers are often framed in standard engineering terms, function is explained using standard engineering concepts. Thus, one reads about mechanisms that 'turn up the gain,' 'act as a resistor,' 'apply a cut-off filter,' 'use a step-function,' 'shift a spectral sensitivity,' and so on. Surely, the objector says (or must be saying), we could profitably recast this 'engineering talk' in terms of detection and indeed surely we shall, once the full explanations are in hand. I think not. Let me give one short example of a sensory process for which this sort of 'recasting,' far from illuminating function, would positively obscure it.

Consider our proprioceptive system, the sense through which we know where our limbs are in space (without looking, that is).[9] In this system, the receptors are not external transducers, ones affected by events beyond or on

the skin.[10] Rather, our sense of proprioception relies upon muscle spindles, internal receptors that are stimulated by the mechanical stretching of the muscles at our joints. In this system, the 'engineering' dilemma that arises is that the range of activity to be encoded exceeds the ability of the neurons to respond. The elbow joint, for example, has a wide range of angular extension; as you extend your arm from its fully bent to the straight position, the flexor and tensor muscles are shortened and lengthened considerably. Sensory neurons, however, have a fairly restricted range of response—there is a limit on just how quickly a neuron can fire (that is, four times a second). One has, then, what looks like a dilemma: if the muscle spindles are set to respond over the full range of joint movement, they will give only coarse-grained information about muscle stretch; but if spindles are adjusted to provide fine-grained information about muscle extension, they will respond over only a limited range of joint movement. What to do? The neural solution to this dilemma is a certain kind of feedback process: central control alters the sensitivity of the receptors, the response of the muscle spindles to change in muscle length. As the limb extends, and as the muscle becomes longer and longer, central control 'lowers the gain' on the spindle response. This allows the spindle to maintain a continuing, fine-grained response throughout the entire range of muscle movement—although it does so at the 'expense' of determining muscle length or extension.

This kind of problem, encountered by the proprioceptive system, is a very common one in sensory processing. Quite often the range of stimulation over which the system must respond far exceeds the ability of a sensory neuron to signal those events. (For example, the visual system of the cat, it is thought, responds over an illumination range of about sixteen log units!) Sometimes the solution to the problem is to build a variety of receptors, in order to cover the full range of stimuli; just as often, the solution is to use a mechanism that adjusts sensitivity. What is important to realize, here, is that there need not be any further device that records the 'position' of the gain mechanism. If the purpose of a stretch receptor is to keep the flexor and tensor muscles balanced (as one stretches, the other contracts), the absolute muscle length is not what is important (at least not for this task). One could say, of course, that the stretch receptors detect muscle length if one wanted to give a rough characterization of what they do (after all, they do react to mechanical stimulation—the muscle stretching). But this way of talking quickly becomes misleading—say, in the face of motor disfunction—if one wants to explain why a person has tensor rigidity (or spasticity). For a careful characterization, one that will explain both function and misfunction, talk of detection only confuses the matter. Thus, while the proximal stimulus is the stretching of the muscle, the function of the muscle spindles is characterized more broadly, in terms of what the system is doing.

As I said above, the third objection depends upon a very strong and broad empirical conjecture about sensory function. Quite obviously, given that neurobiology is just getting off the ground and that, as a result, there are almost no complete accounts of any single sensory system, this question, about the correct principle of sensory processing, remains open. Still, the objector's claim is very strong, and (my guess is) not well grounded in any neurobiological evidence. Is there any other science, then, apart from neurobiology, that might prove the detection thesis?

A different defense of the detection thesis is based upon general evolutionary considerations. After all, the defense goes, organisms simply could not have survived if their senses did not function as reliable detectors of salient properties. If animals are going to find prey, avoid predators, and attract mates, then their sensory systems must detect these properties of the world. As Patricia Churchland[11] once said, in the course of defending an early theory of representational content called *correlational content* 'I take it as obvious that if there were no systematic relations between the external world and states of the brain, the animals could not survive or if they did it would be a miracle in the strict sense of the word. An owl would not know where the mouse is and he would not intercept it as it runs' (*op. cit.*, p. 260). Stating this view in an even stronger form, Dretske says:

> Without such internal indicators, an organism has no way to negotiate its way through its environment, no way to avoid predators, find food, locate mates and do the things it has to do to survive and propagate. ***This, indeed, is what perception is all about*** (italics mine). An animal's senses . . . are merely the diverse ways nature has devised for making what happens inside an animal depend, in some indicator-relevant way, on what happens outside. If the firing of a particular neuron in a female cricket's brain did not indicate the distinctive chirp of a conspecific male, there would be nothing to guide the female in its efforts to find a mate.[12]

Moreover, because evolutionary theory explains just those factors which affect a species' survival or demise, evolutionary biology will make reference to exactly those environmental properties to which the creature reacts (or fails to react). To illustrate, a typical explanation in evolutionary biology might go something like this: 'at the beginning of the egg-laying season, the parasites choose with care their egg-laying sites, rejecting less than optimal places; when the days grew shorter, signaling the end of the season, the parasites became far less discriminating, choosing exactly those kinds of sites which were rejected earlier.' In the standard literature of behavioral ecology and evolutionary biology, even parasites are seen to recognize 'optimal and suboptimal egg-laying sites,' to detect the 'shortening of the days' and 'the end of the breeding season.' What allows us to advance the detection thesis, then, is the simple fact that, in order to survive, all animals must react to salient properties of the environment—and evolutionary biologists cannot explain the factors affecting survival without characterizing the organism's reactions as such.

Again, there is much truth to the view. Of course, it is true that, as a whole, an animal's behavior must be directed toward certain salient objects or properties of its environment. Objects (and their properties) are important to the survival of all creatures. But from this fact alone, one cannot infer that the system of sensory encoding, used to produce that behavior, uses a veridical encoding. That is, it does not follow from the fact that the owl's behavior is *directed toward* the mouse or that the brain states bear systematic relations to stimuli, that there are any states of the owl's sensory system that are *about* or serve to detect that property of being a mouse. What is required for survival of the owl, for example, is that it finds its way to the mouse in a timely fashion and that, once it gets there, the owl can grab the mouse. To do so, the owl's visual system need not encode external space veridically, with anything like, say, a cartographer's topographic map, with the spatial relations of the world exactly mirrored by the spatial relations of the sensory map. Nor need the owl intercept the mouse by literally computing its trajectory based upon its present motion, nor need it use the deliverances of a mouse detector as part of the calculation. It is an open question what kinds of systems are actually used. In the case of the cricket, detector cells, for the chirp of the male conspecific, *are* used. But this is only one neural solution to one behavioral problem. Nothing about the directedness of an organism's *behavior* yields a firm conclusion about the directedness of the *internal states* of its sensory system.

Setting aside the question of whether there is any empirical evidence for the detection thesis, let me turn to a second response to it. Even if we granted the objection—even if we admitted that narcissistic properties will invariably figure in our neurobiological descriptions—this would not give the naturalist what he wants, namely, the beginnings of a theory of aboutness. The suggestion made is that states of sensory systems are about narcissistic properties, properties defined by relation to the subject's interests—*relational properties*. But what the naturalist wishes to explain, in the end, are representations that are about *objective properties* and events of the world. Whatever the neurophysiological facts about the thermoreceptive system, that is, we do not merely *feel* temperature sensations qua sensations in our skin. We have thoughts about the independent temperatures of objects. I put my hand out, grasp the coffee cup, and feel its warmth. I see it *as* a coffee cup and feel the cup *as* having a particular property, warmth. Similarly, when I put my hands into the dishwater, I feel *the water* as being warm, as having an objective property that is independent of my perceptions. Consider again the illusion produced by cooling one hand and warming the other, then placing both hands in a single bucket of tepid water. Here, the illusion, as you place both hands in the tepid water, is that *the water* is both hot and cold. That is what makes the exercise so surprising. There would be no illusion of a contradiction, however, if one assigned content as the objector suggests. Let one sensation indicate the property 'precipitous drop in right-hand temperature' and the other sensation indicate 'precipitous increase in left-hand temperature' and there is no explicit contradiction to be found. So the problem for the naturalist lies in the rather large gap between what our ordinary perceptions are about and the representational contents that the detection of narcissistic properties would assign, namely, mental states that are about 'what is good for me,' subject-dependent properties, as it were. To put this another way, the naturalist's hope was to find a relation between sensory states and external properties, a relation that would ground (in some undefined way) a theory of aboutness. But in order to save the traditional view, the objector introduces narcissistic properties and thereby gives up the link

to the objective properties of the world. Prima facie, this is not a promising starting point for the naturalist theory. I shall have more to say about this later.

This same argument against the third objection, the detection thesis, also works against the first objection, the a priori defense. On the a priori view, recall, sensory mechanisms define properties of the world—large disjunctive sets of *objective properties.* Thus, on the a priori view, an individual temperature sensation defines a property, the disjunctive set of exactly those thermal stimuli which would evoke a particular sensation. The 'hot' feeling will define one 'property,' a disjunctive set of thermal properties; the 'cold' feeling will define another different disjunctive set. Does not this way of describing the properties 'detected,' as disjunctive sets of nonrelational properties, circumvent the problem? No. Even given this redescription of the properties, the a priori view fails to capture the content of ordinary thermal perceptions. When I put both hands into tepid water, the properties detected by the left hand and the right hand comprise two different disjunctive sets. *Ex hypothesis* one temperature of water can bring about these two different properties in the course of normal function. Hence there is no contradiction between what my right hand says and what my left hand says (as there would be if there were no overlap between the disjuncts of both sets). In retrospect, this is exactly the problem one would expect given the form of the a priori defense: it simply takes a set of relational, narcissistic properties and redefines them in nonrelational terms. Defining the very same 'properties' by means of a different device, however, does not alter the very nature of those properties. Thus, even though the a priori views picks out sets of objective properties with which to match our temperature perceptions, it, too, picks out the *wrong* sets.

IV. Philosophical Implications

So far, nothing that has been said casts any doubt upon our common-sense view about our perceptions qua intentional representations of the world. We have thoughts about the temperature of objects. We have thoughts about the Eiffel tower (real or fictive), about George Smith (the very one I met at Tufts), and the fragrance of gardenias (however vague or unreliable). However our sensory systems work, we are creatures who represent a world of objects, properties, and events. Granting all that has been said above, there is no reason, here, to question the existence of mental representations writ large. The philosophical problem of aboutness still stands.

On the other hand, if the above arguments are to believed, sensory systems do not seem to help us understand this fact. Let us take stock of the situation. First, sensory systems exemplify (usually quite elegant) solutions to specific processing problems. Sometimes reliable correlations are used (as in the case of the frog's proverbial fly detector) and sometimes they are not (as in the case of the thermoreceptors). The wiring fits the problem. Second, these kinds of nonrepresentational systems are characteristic of not only simple creatures—for example, the ones that swim around. These are the kinds of systems which connect *us* to the world. At our sensory and motor peripheries, our systems have narcissistic properties. Third, much of a simple organism's behavioral repertoire can be accounted for without the use of anything other than such narcissistic systems, without anything that looks like an internal representation of objects and properties. If one wants to build a simple organism that swims around and eats anchovies, it is not clear that *any* representational states must be used at all. For many classes of behavior, some of which are quite complex, 'making it work' is all that counts. This is how sensory function looks from the point of view of the neurophysiologist.[13]

What lessons about representation can be wrought from these two viewpoints? First, insofar as the neurophysiologist's view conflicts with our firm intuitions about how a sensory system should work, we ought to be suspicious of our intuitions. These are, after all, intuitions about a quintessentially empirical question—'How do sensory systems work'?—a question to which a scientific answer is only now forming. Perhaps we have simply painted sensory systems in our own first-person image. Because our own conscious perceptions are genuinely representational (we have thoughts about the world, its objects, properties, and events), we have expected that both the sensory systems of simple creatures and the parts of our own complex systems will have this property as well, or at least some rough facsimile of it. We have assumed, without warrant, that all sensory systems (or parts thereof) are, in some sense or other, about the world as well.

More importantly, however, the distance between the neurophysiologist's view of sensory

systems and our first-person perspective on conscious perception should raise a genuine puzzle (or rather, yet another genuine puzzle) about representation. We ought to wonder why and how we came to represent the world as we do, given the way in which our neural systems anchor us to the world, given how our sensory and motor systems function. The initial assumption of naturalistic theories has been that even simple perceptual systems must have states that have, at least in a limited sense, aboutness. If an organism did not know, at least roughly, what was out there, when, and where, how could it possibly survive? This is why even simple systems must represent the objects and properties of the world. But if the function of sensory systems is not to inform the brain (the ganglia?) about properties of the world per se, then genuinely representational systems are not merely the next bells and whistles added on to an established evolutionary trend. (First there was representation, and then there was *more*.) They are not mere refinements of or embellishments on an ongoing representational strategy. Our ability to represent the external world as containing objects, properties, and events constitutes a distinct—different—capacity of an organism. What exactly is this capacity, and for what reasons did it come about? If an organism can get about, feed itself, and reproduce all without representational mechanisms, what neural/environmental 'problem' was answered by the evolution of intentional states? How does this representational capacity work and, more specifically, how do our conscious intentional perceptions seem to form an apparently seamless union with our narcissistic sensory systems?

Let me try to set out this puzzle in more detail. On this view of things, aboutness constitutes something like an ontological 'capacity,' an ability to impose stability, order, and uniformity upon a conception of the world (and sometimes the world itself) on the basis of stimuli that do not themselves exhibit these properties. Walk around your kitchen and imagine the images that thereby are reflected on the back of your retinae (pretend, that is, that you are Bishop Berkeley making a cup of tea). The image of your refrigerator looms larger and then smaller as you walk toward it, then back away; despite uniform paint, its three sides have different spectral reflectances (given the location of the light source and each side's proximity to other colored surfaces); as you turn your head, the image shape changes dramatically; when you turn around and walk out the kitchen door, the image disappears completely. Add to this the fact that the images on both retinae are somewhat different, for each eye views the world from a slightly different vantage point. And on and on. That none of these changes in the stimuli matters to your perception of the world is a remarkable fact. You see the fridge as having a constant shape, with a uniform surface color, standing in a single place in the world. Despite the differences between this fridge and the hundreds of others you have seen, you regard it as one instance of a type, as a refrigerator; despite the dents and scratches it has gathered over the past years, you know it is the very same refrigerator that you purchased six years ago, the same one that sat in a warehouse several miles away. That you come to glean this stable ontology, of particulars that instantiate types, of particulars that occupy stable places in the world, is an astounding capacity. It requires that you (your brain) find stability despite stimulus change and uniformity, despite real difference, despite the dissimilarities between objects of the same type and the changes in objects over time. It does not matter whether, in the world, these stabilities and uniformities actually exist (or whether, in some cases, we merely impose stability via ontological categories). To conceive of types and tokens, places and objects as existing at all, given our sensory access to the world, is a fantastically difficult task. Call this the *ontological project*.

On the other hand, there is that small task, assigned to sensory systems, of getting the job done, of directing motor behavior. For the most part, this is not an ontological project, a task for which it is an advantage to see the world according to stable categories or as containing re-identifiable places and particulars. For one, because sensory systems encode information symbiotically with motor needs, the similarities and differences, uniformities and discontinuities that the senses 'record' need not exist in the world. What matters is whether the system of encoding is effective, whether the encoded similarities and differences are useful.

The vestibular system affords an excellent example here. When the vestibular system, which provides our sense of balance, records the tilt of your head away from the vertical, it 'partitions' this information in accordance with the requirements of keeping a human body upright. The physics of falling bodies being what they are, it is much easier to right oneself as the body begins to tilt or lean away from the vertical. So around the vertical point itself, the system makes extremely

fine discriminations, thereby allowing the motor system to make the appropriate fine-grained postural adjustments. As the body tilts farther and farther away from the vertical, coarser discriminations (of angle of tilt) are made. So if one were to draw a polar map of human vestibular function, with the vertical at the center, the map would show concentric rings, with denser rings around the center, with wider and wider spaces between the circles, until the forty-degree mark. (Why forty degrees? After forty degrees, the physics win. The vestibular system ceases to respond because the motor system cannot right the human body after that point.[14]) From your point of view, that of the conscious subject, very large changes in tilt, once you are already off balance, all count as the same, while the fine differences around the vertical which you do acknowledge do not mark any uniform fact about your position. The vestibular system's representations of tilt, its 'concentric rings,' mark imaginary divides that are useful for keeping you upright. In general, sensory systems both make and ignore distinctions in nature when it suits the organism's motor needs.

There is another reason, more difficult to explain, why sensory systems might be considered 'pre-ontological,' or unconcerned with a delineated world. Take the philosopher's favorite reptile, the frog. When the fly-detector cells react, they indicate (in normal circumstances) the presence of a fly, but not the presence of any particular fly (call him 'Herbert'). Just a fly, whichever one happens to present itself. For a frog, at least, it does not matter whether this fly is Herbert or Harold; it does not need to keep track of or identify Herbert *as a particular*, whatever that could mean ('You've flown your last flight, Herbert!'). Nor, for the purposes of eating the fly, does it matter that Herbert is in any particular place. The co-ordinates of the moving spot on the retina encode the fly's position relative to the frog and this is exactly the information which the tongue-swiper needs—the direction of swipe for the tongue relative to the frog. Nor, for the purposes of consuming the fly, does the frog need to know where, in the world, Herbert is located ('The last time I saw Herbert, he was sitting in the Savoy'). All of these are familiar points from the philosophical literature. The same lesson, however, can be applied to our own case. When push comes to shove, so to speak, much of the information needed to make a movement is of the very same sort. Although you see the fly as a particular (and may even call him 'Herbert'), reaching out to grab the fly will require information, not about Herbert qua particular fly (here in the lounge of the Savoy), but about his velocity relative to your arm motion, his position relative to your hand, and so on.[15] *That* information, without which you cannot catch the fly, is no different in kind than the information required by the frog. It is extremely precise information about a particular but not *qua* a particular. Call the narcissistic encoding of this type of information the *sensory-motor project*.

It is the gap between the needs of the sensory-motor project and the demands of the ontological project, I want to claim, that calls out for explanation. What were the behavioral/environmental conditions in virtue of which the development of an ontology—and hence things and properties to have thoughts about—came to have survival value to our predecessors?[16] Exactly what kinds of abilities or capacities are required in order to represent a stable ontology, of types and tokens, objects and places? And how exactly does the information provided by our sensory systems co-exist with, form a whole with, the ontology imposed by a representational system? (Given that we *feel* thermal sensations as a function of a receptor population response, how does that fit with our conception of temperature as an objective property of objects?) These are the questions about representational directedness which immediately arise.

To grant that there is this sort of puzzle—even to ask the above questions about directedness without answering them—is to admit, pace Wilfrid Sellars, that in an important sense we do not really know what 'aboutness' is. Certainly, at the outset, a vague realism about the directedness of mental/neural events is adopted: representations are 'tied' to objects and properties and hence (there being no good reason to suppose otherwise) bear some kind of relation to them. But if we do not know exactly what it means to regard a particular as a particular, to see this thing as being of a certain type, this place as the same place, and so on—hence what kinds of capacities or abilities are involved in having representations that are about those things—then we do not know, in any substantive sense, in what that relationship consists. We only trust that it *is*.

To admit the above, I think, is to take on a different view about what kind of intertheoretic explanations are appropriate to intentional phenomena. On the naturalistic scheme, recall, the idea is to explain the relation of aboutness, in the first instance, by appeal to some other relation

postulated by the natural sciences. Exactly how this natural relation figures in the larger theoretical picture varies from theory to theory, given the diversity of the authors' explanatory goals. Sometimes the natural relation is thought to confer a sort of 'proto-intentionality' upon specific sensory states, states that are hypothesized to form a 'ground' for the genuinely intentional states of folk theory (see, for example, Dretske's *Knowledge and the Flow of Information*); in other theories, the natural relation is simply identified with the 'aboutness' relation of, say, a computational state (for example, the Churchland's notion of 'calibrational content,' developed in Paul Churchland's *Scientific Realism and the Plasticity of Mind*). Once one realizes that the demands of the sensory-motor project are, for the most part, distinct from the demands of the ontological project, however, one realizes that sensory systems need not be veridical reporters as portrayed by the traditional view. Hence sensory states need not bear the expected natural relation to external properties—the natural relation that was to mirror or ground the aboutness of mental or computational states. To ask the above questions is to admit, then, that the directedness of mental events constitutes a distinct set of representational abilities and capacities, which at this point, we can only roughly define. If this is so, then there is no longer any reason to think that a relational property at one level of theoretic explanation (the psychological/computational level) will have a clear mapping onto, or grounding in, any relational property at another (the neurophysiological or biological). The relation of aboutness need not be explained primarily in terms of some other natural relation at all.

Most importantly, questions of the above kind serve to shift the focus of theoretic attention away from static perceptual states. Recall the naturalist's hope that by closely scrutinizing simple sensory events and their relations to external causes, we would gain a toehold on the phenomenon of directedness—by understanding the most straightforward cases of mental directedness, we would have a route into more complex intentional phenomena. Because all of the senses, on the traditional view, are veridical, it will be the static perceptual case qua correlational state that will provide the essential key to aboutness. Once the assumptions of the traditional view are set aside, however, there is no assumption that sensory states will, in general, be about external properties of the world. More importantly, one can no longer assume that for those fully intentional perceptual states which indeed are about the world, it is the static perceptual case—where we sit with our eyes open staring at the object of perception—that will provide us with insight into that relation. Trace out the causal path between the object of perception, the stimulation of the receptors, and whatever neural events that thereafter eventuate and this alone will not *explain*, in the required sense, how genuine representation arises. The explanation of any particular perception can take place only against a background theory of the representational capacities at work. So, it is a theoretic understanding of those capacities qua capacities which will give the explanation of aboutness bite—and these are not capacities which we will understand simply by scrutinizing with care the static perceptual case. (This is not to say that the study of sensory systems is of no utility to the intentionality theorist—on the contrary—but that looking for correlations between sensory states and external properties is not what the study of perception will be all about.)

NOTES

1. See, for example, Paul Churchland's *Scientific Realism and the Plasticity of Mind* (New York: Cambridge, 1979); a joint article by both Paul and Patricia Churchland, "Functionalism, Qualia and Intentionality," *Philosophical Topics* I (Spring 1981): pp. 121–45; Patricia Churchland's *Neurophilosophy: Toward a Unified Science of the Mind-Brain* (Cambridge, MA: MIT Press, 1986); Dennett's *The Intentional Stance* (Cambridge, MA: MIT Press, 1987); Dretske's *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981) and *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988); Fodor's *Psycho semantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press, 1987) and *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990); Papineau's "Representation and Explanation," *Philosophy of Science* LI, no. 4 (December 1984): pp. 550–72; Stampe's "Show and Tell," in *Forms of Representation*, B. Freeman et al, eds., (Amsterdam: North-Holland, 1975), pp. 221–45; and Sterelny's *The Representational Theory of Mind* (Cambridge, MA: MIT Press, 1990).
2. Of course, simplicity is not the only reason why the static perceptual case has seemed such an obvious starting point. Nor need this assumption hinge entirely on any view of the senses. For philosophers with an empiricist bent, who believe that the content of ***all*** of our mental/computational states is somehow wrought from the content of perceptual states, perception is not merely the most simple case of

mental directedness. Perceptual states are also the most ***basic***. If it is our perceptions that somehow provide the raw materials for our more complex intentional states (about unperceived, fictive, or abstract properties), then perceptual states ***ground*** all other intentional relations. For an empiricist, this methodological point of entry is dictated by his philosophical program. Even a rationalist, however, will find the study of the simple perceptual case a perfectly obvious starting point. All one need believe is that, in the absence of all sensory input, there would not be (could not be) any thought at all, that sensory input is a necessary condition for intentional states. On this view, perceptual states are also 'basic,' although not in the strong sense endorsed by the empiricist, given above.

3. For a well-worked out example of the notion of isomorphic representations, see C.R. Gallistel's *The Organization of Learning* (Cambridge, MA: MIT Press, 1990).
4. *Thermal Sensations and Thermoreceptors in Man* (Springfield, IL: Thomas, 1982).
5. To make this point more tangible, try to imagine the thermoreceptive system of some other creature—say, the penguin.
6. For a very similar view of sensory processing, see Patricia Churchland, V. S. Ramahandran, and Terry Sejnowski, "A Critique of Pure Vision," in *Large-Scale Neuronal Theories of the Brain*, C. Koch and J. L. Davies, eds., (Cambridge, MA: MIT Press, 1994), pp. 23–60.
7. In *Consciousness Explained* (New York: Little, Brown, 1991), p. 382, Dennett says: 'What property does Otto judge something to have when he judges it to be pink? The property he calls 'pink.' And what property is that? It's hard to say but that should not embarrass us, because we can say why it is hard to say. The best we can do, practically, when asked what surface properties we detect with color vision, is to say, uninformatively, that we detect the properties we detect. If someone wants a more informative story about these properties there is a large and incompressible literature in biology, neuroscience, and psychophysics to consult.'
8. My thanks here to José Antonio Diez Calzada for clarifying this point.
9. For an extremely interesting account of what it is like to live ***without*** proprioception, that is, with only visual knowledge of body position, see Jonathan Cole's *Pride and a Daily Marathon* (Cambridge, MA: MIT Press, 1995).
10. The following information on stretch receptors comes from T. Carew and C. Ghez, "Muscles and Muscle Receptors," in *Principles of Neural Science*, Eric Kandel and James Schwartz, eds., (New York: Elsevier, 1985), 2nd ed., pp. 443–56, here p. 454 ff.
11. "Replies to Comments" from "Symposium on Patricia Smith Churchland's Neurophilosophy," *Inquiry* XXIX (June 1986): pp. 139–273.
12. *Explaining Behavior: Reasons in a World of Causes*, p. 62.
13. And to many computer scientists. For one defense of this view, see Rodney Brook's "Intelligence without Representations," *Artificial Intelligence* XLVII (January 1991): pp. 139–60.
14. And why, one might ask, are cats so good landing on their feet? Because, through a neck reflex, they manage to keep their heads, and vestibular systems, turned upright thereby receiving vestibular information through almost the entire range of body angles. That, of course, and faster righting reflexes in general.
15. For some interesting research on cortical visual maps that are 'arm-centered,' see Micheal Granziano, Gregory Yap, and Charles Gross's "Coding of Visual Space by Premotor Neurons," *Science*, CCLXVI (November 1994): pp. 1054–57.
16. To put this another way, there are certain circumstances when particulars (one's off-spring or mate) impinge upon the lives of simple creatures and for which having detector cells can be just what is needed. Detectors are reliable (indeed, they are often much more reliable than, say, the mechanisms of human facial recognition) and they are usually economical as well, requiring few neural resources (an ant, one must realize, does not have many neurons to spare). But while detectors may signal the presence of a particular, they do not represent them ***as*** particulars. So one can rephrase the above question as follows: What kinds of informational problems about particulars require, or are better served by, something other than detectors?

Biosemantics

Ruth Garrett Millikan

Causal or informational theories of the semantic content of mental states which have had an eye on the problem of false representations have characteristically begun with something like this intuition. There are some circumstances under which an inner representation has its represented as a necessary and/or sufficient cause or condition of production. That is how

From *Journal of Philosophy* 86 (1989): pp. 281–97. Reprinted with permission of author and publisher.

the content of the representation is fixed. False representations are to be explained as tokens that are produced under other circumstances. The challenge, then, is to tell what defines certain circumstances as the content-fixing ones.

I

Note that the answer cannot be just that these circumstances are *statistically* normal conditions. To gather such statistics, one would need to delimit a reference class of occasions, know how to count its members, and specify description categories. It would not do, for example, just to average over conditions-in-the-universe-any-place-any-time. Nor is it given how to carve out relevant description categories for conditions on occasions. Is it 'average' in the summer for it to be (precisely) between 80° and 80.5° Fahrenheit with humidity 87%? And are average conditions those which obtain on at least 50% of the occasions, or is it 90%? Depending on how one sets these parameters, radically different conditions are 'statistically normal.' But the notion of semantic content clearly is not relative, in this manner, to arbitrary parameters. The content-fixing circumstances must be *nonarbitrarily* determined.

A number of recent writers have made an appeal to teleology here, specifically to conditions of normal function or well-functioning of the systems that produce inner representations. Where the represented is *R* and its representation is 'R,' under conditions of well-functioning, we might suppose, only *R*s can or are likely to produce 'Rs.' Or perhaps 'R' is a representation of *R* just in case the system was designed to react to *R*s by producing 'Rs.' But this sort of move yields too many representations. Every state of every functional system has normal causes, things that it is a response to in accordance with design. These causes may be proximate or remote, and many are disjunctive. Thus, a proximate normal cause of dilation of the skin capillaries is certain substances in the blood, more remote causes include muscular effort, sunburn, and being in an overheated environment. To each of these causes the vascular system responds by design, yet the response (a red face), though it may be a natural sign of burn or exertion or overheating, certainly is not a representation of that. If not every state of a system represents its normal causes, which are the states that do?

Jerry Fodor[1] has said that, whereas the content of an inner representation is determined by some sort of causal story, its status *as* a representation is determined by the functional organization of the part of the system which uses it. There is such a thing, it seems, as behaving like a representation without behaving like a representation of anything in particular. What the thing is a representation of is then determined by its cause under content-fixing conditions. It would be interesting to have the character of universal I-am-a-representation behavior spelled out for us. Yet, as Fodor well knows, there would still be the problem of demonstrating that there was only one normal cause per representation type.

A number of writers, including Dennis Stampe,[2] Fred Dretske,[3] and Mohan Matthen,[4] have suggested that what is different about effects that are representations is that their function is, precisely, to represent, 'indicate,' or 'detect.' For example, Matthen says of (full-fledged) perceptual states that they are 'state[s] that [have]the function of *detecting* the presence of things of a certain type . . .' (ibid., p. 20). It does not help to be told that inner representations are things that have representing (indicating, detecting) as their function, however, unless we are also told what kind of activity representing (indicating, detecting) is. Matthen does not tell us how to naturalize the notion 'detecting.' If 'detecting' is a function of a representational state, it must be something that the state effects or produces. For example, it cannot be the function of a state to have been produced in response to something. Or does Matthen mean that it is not the representational states themselves, but the part of the system which produces them, which has the function of detecting? It has the function, say, of producing states that correspond to or covary with something in the outside world? But, unfortunately, not every device whose job description includes producing items that vary with the world is a representation producer. The devices in me that produce calluses are supposed to vary their placement according to where the friction is, but calluses are not representations. The pigment arrangers in the skin of a chameleon, the function of which is to vary the chameleon's color with what it sits on, are not representation producers.

Stampe and Dretske do address the question what representing or (Dretske) 'detecting' is. Each brings in his own description of what a natural sign or natural representation is, then assimilates *having the function of representing R* to being a natural sign or representer of *R* when

the system functions normally. Now, the production of natural signs is undoubtedly an accidental side effect of normal operation of many systems. From my red face you can tell that either I have been exerting myself, or I have been in the heat, or I am burned. But the production of an accidental side effect, no matter how regular, is not one of a system's functions; that goes by definition. More damaging, however, it simply is not true that representations must carry natural information. Consider the signals with which various animals signal danger. Nature knows that it is better to err on the side of caution, and it is likely that many of these signs occur more often in the absence than in the presence of any real danger. Certainly there is nothing incoherent in the idea that this might be so, hence that many of these signals do not carry natural information concerning the dangers they signal.

II

I fully agree, however, that an appeal to teleology, to function, is what is needed to fly a naturalist theory of content. Moreover, what makes a thing into an inner representation is, near enough, that its function is to represent. But, I shall argue, the way to unpack this insight is to focus on representation *consumption*, rather than representation production. It is the devices that *use* representations which determine these to be representations and, at the same time (contra Fodor), determine their content. If it really is the function of an inner representation to indicate its represented, clearly it is not just a natural sign, a sign that you or I looking on might interpret. It must be one that functions as a sign or representation *for the system itself.* What is it then for a system to use a representation *as* a representation?

The conception of function on which I shall rely was defined in my *Language, Thought, and Other Biological Categories*[55] and defended in "In Defense of Proper Functions"[6] under the label 'proper function.' Proper functions are determined by the histories of the items possessing them; functions that were 'selected for' are paradigm cases.[7] The notions 'function' and 'design' should not be read, however, as referring only to origin. Natural selection does not slack after the emergence of a structure but actively preserves it by acting against the later emergence of less fit structures. And structures can be preserved due to performance of new functions unrelated to the forces that originally shaped them. Such functions are 'proper functions,' too, and are 'performed in accordance with design.'

The notion 'design' should not be read—and this is very important—as a reference to innateness. A system may have been designed to be altered by its experience, perhaps to learn from its experience in a prescribed manner. Doing what it has learned to do in this manner is then 'behaving in accordance with design' or 'functioning properly.'[8]

My term 'normal' should be read normatively, historically, and relative to specific function. In the first instance, 'normal' applies to explanations. A 'normal explanation' explains the performance of a particular function, telling how it was (typically) historically performed on those (perhaps rare) occasions when it was properly performed. Normal explanations do not tell, say, why it has been common for a function to be performed; they are not statistical explanations. They cover only past times of actual performance, showing how these performances were entailed by natural law, given certain conditions, coupled with the dispositions and structures of the relevant functional devices.[9] In the second instance, 'normal' applies to conditions. A 'normal condition for performance of a function' is a condition, the presence of which must be mentioned in giving a full normal explanation for performance of that function. Other functions of the same organism or system may have other normal conditions. For example, normal conditions for discriminating colors are not the same as normal conditions for discriminating tastes, and normal conditions for seeing very large objects are not the same as for seeing very small ones. It follows that 'normal conditions' must not be read as having anything to do with what is typical or average or even, in many cases, at all common. First, many functions are performed only rarely. For example, very few wild seeds land in conditions normal for their growth and development, and the protective colorings of caterpillars seldom actually succeed in preventing them from being eaten. Indeed, normal conditions might almost better be called 'historically optimal' conditions. (If normal conditions for proper functioning, hence survival and proliferation, were a statistical norm, imagine how many rabbits there would be in the world.) Second, many proper functions only need to be performed under rare conditions. Consider, for example, the vomiting reflex, the function of which is to prevent

(further) toxification of the body. A normal condition for performance of this function is presence, specifically of poison in the stomach, for (I am guessing) it is only under that condition that this reflex has historically had beneficial effects. But poison in the stomach certainly is not an average condition. (Nor, of course, is it a normal condition for other functions of the digestive system.[10])

If it is actually one of a system's functions to produce representations, as we have said, these representations must function as representations for the system itself. Let us view the system, then, as divided into two parts or two aspects, one of which produces representations for the other to consume. What we need to look at is the consumer part, at what it is to use a thing *as* a representation. Indeed, a good look at the consumer part of the system ought to be all that is needed to determine not only representational status but representational content. We argue this as follows. First, the part of the system which consumes representations must understand the representations proffered to it. Suppose, for example, that there were abundant 'natural information' (in Dretske's[11] sense) contained in numerous natural signs all present in a certain state of a system. This information could still not serve the system *as* information, unless the signs were understood by the system, and, furthermore, understood as bearers of whatever specific information they, in fact, do bear. (Contrast Fodor's notion that something could function like a representation without functioning like a representation of anything in particular.) So there must be something about the consumer that *constitutes* its taking the signs to indicate, say, *p*, *q*, and *r* rather than *s*, *t*, and *u*. But, if we know what constitutes the consumer's *taking* a sign to indicate *p*, what *q*, what *r*, etc., then, granted that the consumer's takings are in some way systematically derived from the structures of the signs so taken, we can construct a semantics for the consumer's language. Anything the signs may indicate qua natural signs or natural information carriers then drops out as entirely irrelevant; the representation-producing side of the system had better pay undivided attention to the language of its consumer. The sign producer's function will be to produce signs that are true *as the consumer reads the language.*

The problem for the naturalist bent on describing intentionality, then, does not concern representation production at all. Although a representation always is something that is produced by a system whose proper function is to make that representation correspond by rule to the world, what the rule of correspondence is, what gives definition to this function, is determined entirely by the representation's consumers.

For a system to use an inner item as a representation, I propose, is for the following two conditions to be met. First, unless the representation accords, *so* (by a certain rule), with a represented, the consumer's normal use of, or response to, the representation will not be able to fulfill all of the consumer's proper functions in so responding—not, at least, in accordance with a normal explanation. (Of course, it might still fulfill these functions by freak accident, but not in the historically normal way.) Putting this more formally, that the representation and the represented accord with one another, so, is a normal condition for proper functioning of the consumer device as it reacts to the representation.[12] Note that the proposal is not that the content of the representation rests on the function of the representation or of the consumer, on what these do. The idea is not that there is such a thing as behaving like a representation of *X* or as being treated like a representation of *X*. The content hangs only on there being a certain condition that would be *normal* for performance of the consumer's functions—namely, that a certain correspondence relation hold between sign and world—whatever those functions may happen to be. For example, suppose the semantic rules for my belief representations are determined by the fact that belief tokens in me will aid the devices that use them to perform certain of their tasks in accordance with a normal explanation for success only under the condition that the forms or 'shapes' of these belief tokens correspond, in accordance with said rules, to conditions in the world. Just what these user tasks are need not be mentioned.[13]

Second, represented conditions are conditions that vary, depending on the *form* of the representation, in accordance with specifiable correspondence rules that give the semantics for the relevant *system* of representation. More precisely, representations always admit of significant transformations (in the mathematical sense), which accord with transformations of their corresponding representeds, thus displaying significant articulation into variant and invariant aspects. If an item considered as compounded of certain variant and invariant aspects can be said to be 'composed' of these, then we can also say that every representation

is, as such, a member of a representational system having a 'compositional semantics.' For it is not that the represented condition is itself a normal condition for proper operation of the representation consumer. A certain correspondence between the representation and the world is what is normal. Coordinately, there is no such thing as a representation consumer that can understand only one representation. There are always other representations, composed other ways, saying other things, which it could have understood as well, in accordance with the same principles of operation. A couple of very elementary examples should make this clear.[14]

First, consider beavers, who splash the water smartly with their tails to signal danger. This instinctive behavior has the function of causing other beavers to take cover. The splash means danger, because only when it corresponds to danger does the instinctive response to the splash on the part of the interpreter beavers, the consumers, serve a purpose. If there is no danger present, the interpreter beavers interrupt their activities uselessly. Hence, that the splash corresponds to danger is a normal condition for proper functioning of the interpreter beavers' instinctive reaction to the splash. (It does not follow, of course, that it is a usual condition. Beavers being skittish, most beaver splashes possibly occur in response to things not in fact endangering the beaver.) In the beaver splash semantic system, the time and place of the splash varies with, 'corresponds to,' the time and place of danger. The representation is articulate: properly speaking, it is not a splash but a splash-at-a-time-and-a-place. Other representations in the same system, splashes at other times and places, indicate other danger locations.

Second, consider honey bees, which perform 'dances' to indicate the location of sources of nectar they have discovered. Variations in the tempo of the dance and in the angle of its long axis vary with the distance and direction of the nectar. The interpreter mechanisms in the watching bees—these are the representation consumers—will not perform their full proper functions of aiding the process of nectar collection in accordance with a normal explanation, unless the location of nectar corresponds correctly to the dance. So, the dances are representations of the location of nectar. The full representation here is a dance-at-a-time-in-a-place-at-a-tempo-with-an-orientation.

Notice that, on this account, it is not necessary to assume that most representations are true. Many biological devices perform their proper functions not on the average, but just often enough. The protective coloring of the juveniles of many animal species, for example, is an adaptation passed on because *occasionally* it prevents a juvenile from being eaten, though most of the juveniles of these species get eaten anyway. Similarly, it is conceivable that the devices that fix human beliefs fix true ones not on the average, but just often enough. If the true beliefs are functional and the false beliefs are, for the most part, no worse than having an empty mind, then even very fallible belief-fixing devices might be better than no belief-fixing devices at all. These devices might even be, in a sense, 'designed to deliver some falsehoods.' Perhaps, given the difficulty of designing highly accurate belief-fixing mechanisms, it is actually advantageous to fix too many beliefs, letting some of these be false, rather than fix too few beliefs. Coordinately, perhaps our belief-consuming mechanisms are carefully designed to tolerate a large proportion of false beliefs. It would not follow, of course, that the belief consumers are designed to *use* false beliefs, certainly not that false beliefs can serve all of the functions that true ones can. Indeed, surely if none of the mechanisms that used beliefs ever cared at all how or whether these beliefs corresponded to anything in the world, beliefs would not be functioning as representations, but in some other capacity.

Shifting our focus from producing devices to consuming devices in our search for naturalized semantic content is important. But the shift from the *function* of consumers to *normal conditions* for proper operation is equally important. Matthen, for example, characterizes what he calls a 'quasi-perceptual state' as, roughly, one whose job is to cause the system to do what it must do to perform its function, given that it is in certain circumstances, which are what it represents. Matthen is thus looking pretty squarely at the representation consumers, but at what it is the representation's job to get these consumers to do, rather than at normal conditions for their proper operation. As a result, Matthen now retreats. The description he has given of quasi-perceptual states, he says, cannot cover 'real perception such as that which we humans experience. Quite simply, there is no such thing as *the* proper response, or even a range of functionally appropriate responses, to what perception tells us' (op. cit., p. 20).[15] On the contrary, representational content rests not on univocity of consumer function but on sameness of normal conditions for those functions. The

same percept of the world may be used to guide any of very many and diverse activities, practical or theoretical. What stays the same is that the percept must correspond to environmental configurations in accordance with the same correspondence rules for each of these activities. For example, if the position of the chair in the room does not correspond, so, to my visual representation of its position, that will hinder me equally in my attempts to avoid the chair when passing through the room, to move the chair, to sit in it, to remove the cat from it, to make judgments about it, etc. Similarly, my belief that New York is large may be turned to any of diverse purposes, but those which require it to be a *representation* require also that New York indeed be large if these purposes are to succeed in accordance with a normal explanation for functioning of my cognitive systems.

III

We have just cleanly bypassed the whole genre of causal/informational accounts of mental content. To illustrate this, we consider an example of Dretske's. Dretske tells of a certain species of northern hemisphere bacteria which orient themselves away from toxic oxygen-rich surface water by attending to their magnetosomes, tiny inner magnets, which pull toward the magnetic north pole, hence pull down (ibid.). (Southern hemisphere bacteria have their magnetosomes reversed.) The function of the magnetosome thus appears to be to effect that the bacterium moves into oxygen-free water. Correlatively, intuition tells us that what the pull of the magnetosome represents is the whereabouts of oxygen-free water. The direction of oxygen-free water is not, however, a factor in *causing* the direction of pull of the magnetosome. And the most reliable natural information that the magnetosome carries is surely not about oxygen-free water but about distal and proximal causes of the pull, about the direction of geomagnetic or better, just plain magnetic, north. One can, after all, easily deflect the magnetosome away from the direction of lesser oxygen merely by holding a bar magnet overhead. Moreover, it is surely a function of the magnetosome to respond to that magnetic field, that is part of its normal mechanism of operation, whereas responding to oxygen density is not. None of this makes any sense on a causal or informational approach.

But on the biosemantic theory it does make sense. What the magnetosome represents is only what its *consumers* require that it correspond to in order to perform *their* tasks. Ignore, then, how the representation (a pull-in-a-direction-at-a-time) is normally produced. Concentrate, instead, on how the systems that react to the representation work, on what these systems need in order to do their job. What they need is only that the pull be in the direction of oxygen-free water at the time. For example, they care not at all how it came about that the pull is in that direction; the magnetosome that points toward oxygen-free water quite by accident and not in accordance with any normal explanation will do just as well as one that points that way for the normal reasons. (As Socrates concedes in the *Meno*, true opinion is just as good as knowledge so long as it stays put.) What the magnetosome represents then is univocal; it represents only the direction of oxygen-free water. For that is the only thing that corresponds (by a compositional rule) to it, the absence of which would matter—the absence of which would disrupt the function of those mechanisms which rely on the magnetosome for guidance.

It is worth noting that what is represented by the magnetosome is not proximal but distal; no proximal stimulus is represented at all. Nor, of course, does the bacterium perform an inference from the existence of the proximal stimulus (the magnetic field) to the existence of the represented. These are good results for a theory of content to have, for otherwise one needs to introduce a derivative theory of content for mental representations that do not refer, say, to sensory stimulations, and also a foundationalist account of belief fixation. Note also that, on the present view, representations manufactured in identical ways by different species of animal might have different contents. Thus, a certain kind of small swift image on the toad's retina, manufactured by his eye lens, represents a bug, for that is what it must correspond to if the reflex it (invariably) triggers is to perform its proper functions normally, while exactly the same kind of small swift image on the retina of a male hoverfly, manufactured, let us suppose, by a nearly identical lens, represents a passing female hoverfly, for that is what it must correspond to if the female-chasing reflex it (invariably) triggers is to perform its proper functions normally. Turning the coin over, representations with the same content may be normally manufactured in a diversity of ways, even in the same species. How many different ways do you have, for example, of telling a lemon or your spouse? Nor is it necessary that any of the ways

one has of manufacturing a given representation be especially reliable ways in order for the representation to have determinate content. These various results cut the biosemantic approach off from all varieties of verificationism and foundationalism with a clean, sharp knife.

IV

But perhaps it will be thought that belief fixation and consumption are not biologically proper activities, hence that there are no normal explanations, in our defined sense, for proper performances of human beliefs. Unlike bee dances, which are all variations on the same simple theme, beliefs in dinosaurs, in quarks, and in the instability of the dollar are recent, novel, and innumerably diverse, as are their possible uses. How could there be anything *biologically* normal or abnormal about the details of the consumption of such beliefs?

But what an organism does in accordance with evolutionary design can be very novel and surprising, for the more complex of nature's creatures are designed to learn. Unlike evolutionary adaptation, learning is not accomplished by *random* generate-and-test procedures. Even when learning involves trial and error (probably the exception rather than the rule), there are principles in accordance with which responses are selected by the system to try, and there are specific principles of generalization and discrimination, etc., which have been built into the system by natural selection. How these principles normally work, that is, how they work given normal (i.e., historically optimal) environments, to produce changes in the learner's nervous system which will effect the furthering of ends of the system has, of course, an explanation—the normal explanation for proper performance of the learning mechanism and of the states of the nervous system it produces.

Using a worn-out comparison, there is an infinity of functions which a modern computer mainframe is capable of performing, depending upon its input and on the program it is running. Each of these things it can do, so long as it is not damaged or broken, 'in accordance with design,' and to each of these capacities there corresponds an explanation of how it would be activated or fulfilled normally. The human's mainframe takes, roughly, stimulations of the afferent nerves as input, both to program and to run it.[16] It responds, in part, by developing concepts, by acquiring beliefs and desires in accordance with these concepts, by engaging in practical inference leading ultimately to action. Each of these activities may, of course, involve circumscribed sorts of trial and error learning. When conditions are optimal, all this aids survival and proliferation in accordance with an historically normal explanation—one of high generality, of course. When conditions are not optimal, it may yield, among other things, empty or confused concepts, biologically useless desires, and false beliefs. But, even when the desires are biologically useless (though probably not when the concepts expressed in them are empty or confused), there are still biologically normal ways for them to get fulfilled, the most obvious of which require reliance on true beliefs.[17]

Yet how do we know that our contemporary ways of forming concepts, desires, and beliefs do occur in accordance with evolutionary design? Fodor, for example, is ready with the labels 'pop Darwinism' and 'naive adaptationism' to abuse anyone who supposes that our cognitive systems were actually selected for their belief and desire using capacities.[18] Clearly, to believe that every structure must have a function would be naive. Nor is it wise uncritically to adopt hypotheses about the functions of structures when these functions are obscure. It does not follow that we should balk at the sort of adaptationist who, having found a highly complex structure that quite evidently is currently and effectively performing a highly complex and obviously indispensable function, then concludes, *ceteris paribus*, that this function has been the most recent historical task stabilizing the structure. To suspect that the brain has not been preserved for thinking with or that the eye has not been preserved for seeing with—to suspect this, moreover, in the absence of any alternative hypotheses about causes of the stability of these structures—would be totally irresponsible. Consider: nearly every human behavior is bound up with intentional action. Are we really to suppose that the degree to which our behaviors help to fulfill intentions, and the degree to which intentions result from logically related desires plus beliefs, is a sheer coincidence—that these patterns are irrelevant to survival and proliferation or, though relevant, have had no stabilizing effect on the gene pool? But the only alternative to biological design, in our sense of 'design,' is sheer coincidence, freak accident—unless there is a ghost running the machine!

Indeed, it is reasonable to suppose that the brain structures we have recently been using in

developing space technology and elementary particle physics have been operating in accordance with the very same general principles as when prehistoric man used them for more primitive ventures. They are no more performing new and different functions or operating in accordance with new and different principles nowadays than are the eyes when what they see is television screens and space shuttles. Compare: the wheel was invented for the purpose of rolling ox carts, and did not come into its own (pulleys, gears, etc.) for several thousand years thereafter, during the industrial revolution. Similarly, it is reasonable that the cognitive structures with which man is endowed were originally nature's solution to some very simple demands made by man's evolutionary niche. But the solution nature stumbled on was elegant, supremely general, and powerful, indeed; I believe it was a solution that cut to the very bone of the ontological structure of the world. That solution involved the introduction of representations, inner and/or outer, having a subject/predicate structure, and subject to a negation transformation. (Why I believe that that particular development was so radical and so powerful has been explained in depth in *Language, Thought and Other Biological Categories* (LTOBC), chapters 14–19. But see also section v. 6.)

V

One last worry about our sort of position is voiced by Daniel Dennett[19] and discussed at length by Fodor.[20] Is it really plausible that bacteria and paramecia, or even birds and bees, have inner representations in the same sense that we do? Am I really prepared to say that these creatures, too, have mental states, that they think? I am not prepared to say that. On the contrary, the representations that they have must differ from human beliefs in at least six very fundamental ways.[21]

(1) Self-Representing Elements

The representations that the magnetosome produces have three significant variables, each of which refers to itself. The time of the pull refers to the time of the oxygen-free water, the locale of the pull refers to the locale of the oxygen-free water, and the direction of pull refers to the direction of oxygen-free water. The beaver's splash has two self-referring variables: a splash at a certain time and place indicates that there is danger at that same time and place. (There is nothing necessary about this. It might have meant that there would be danger at the nearest beaver dam in five minutes.) Compare the standard color coding on the outsides of colored markers: each color stands for itself. True, it may be that sophisticated indexical representations such as percepts and indexical beliefs also have their time or place or both as significant self-representing elements, but they also have other significant variables that are not self-representing. The magnetosome does not.

(2) Storing Representations

Any representation the time or place of which is a significant variable obviously cannot be stored away, carried about with the organism for use on future occasions. Most beliefs are representations that can be stored away. Clearly this is an important difference.

(3) Indicative and Imperative Representations

The theory I have sketched here of the content of inner representations applies only to indicative representations, representations which are supposed to be determined by the facts, which tell what is the case. It does not apply to imperative representations, representations which are supposed to determine the facts, which tell the interpreter what to do. Neither do causal-informational theories of content apply to the contents of imperative representations. True, some philosophers seem to have assumed that having defined the content of various mental symbols by reference to what causes them to enter the 'belief box,' then when one finds these same symbols in, say, the 'desire box' or the 'intention box,' one already knows what they mean. But how do we know that the desire box or the intention box use the same representational system as the belief box? To answer that question we would have to know what constitutes a desire box's or an intention box's using one representational system rather than another which, turned around, is the very question at issue. In LTOBC and "Thoughts without Laws; Cognitive Science with Content,"[22] I developed a parallel theory of the content of imperative representations. Very roughly, one of the proper functions of the consumer system for an imperative representation is to help *produce* a correspondence between the representation and

the world. (Of course, this proper function often is not performed.) I also argued that desires and intentions are imperative representations.

Consider, then, the beaver's splash. It tells that there is danger here now. Or why not say, instead, that it tells other nearby beavers what to do now, namely, to seek cover? Consider the magnetosome. It tells which is the direction of oxygen-free water. Or why not say, instead, that it tells the bacterium which way to go? Simple animal signals are invariably both indicative and imperative. Even the dance of the honey bee, which is certainly no simple signal, is both indicative and imperative. It tells the worker bees where the nectar is; equally, it tells them where to go. The step from these primitive representations to human beliefs is an enormous one, for it involves the separation of indicative from imperative functions of the representational system. Representations that are undifferentiated between indicative and imperative connect states of affairs directly to actions, to specific things to be done in the face of those states of affairs. Human beliefs are not tied directly to actions. Unless combined with appropriate desires, human beliefs are impotent. And human desires are equally impotent unless combined with suitable beliefs.[23]

(4) Inference

As indicative and imperative functions are separated in the central inner representational systems of humans, they need to be reintegrated. Thus, humans engage in practical inference, combining beliefs and desires in novel ways to yield first intentions and then action. Humans also combine beliefs with beliefs to yield new beliefs. Surely nothing remotely like this takes place inside the bacterium.

(5) Acts of Identifying

Mediate inferences always turn on something like a middle term, which must have the same representational value in both premises for the inference to go through. Indeed, the representation consumers in us perform many functions that require them to use two or more overlapping representations together, and in such a manner that, unless the represenateds corresponding to these indeed have a common element, these functions will not be properly performed. Put informally, the consumer device *takes* these represented elements to be the same, thus identifying their representational values. Suppose, for example, that you intend to speak to Henry about something. In order to carry out this intention you must, when the time comes, be able to recognize Henry in perception as the person to whom you intend to speak. You must identify Henry as represented in perception with Henry as represented in your intention. Activities that involve the coordinated use of representations from different sensory modalities, as in the case of eye-hand coordination, visual-tactile coordination, also require that certain objects, contours, places, or directions, etc., be identified as the same through the two modalities. Now, the foundation upon which modern representational theories of thought are built depends upon a denial that what is thought of is ever placed before a naked mind. Clearly, we can never know what an inner representation represents by a direct comparison of representation to represented. Rather, acts of identifying are our ways of 'knowing what our representations represent.' The bacterium is quite incapable of knowing, in this sense, what its representations are about. This might be a reason to say that it does not understand its own representations, not really.

(6) Negation and Propositional Content

The representational system to which the magnetosome pull belongs does not contain negation. Indeed, it does not even contain contrary representations, for the magnetosome cannot pull in two directions at once. Similarly, if two beavers splash at different times or places, or if two bees dance different dances at the same time, it may well be that there is indeed beaver danger two times or two places and that there is indeed nectar in two different locations.[24] Without contrariety, no conflict, of course and more specifically, no contradiction. If the law of noncontradiction plays as significant a role in the development of human concepts and knowledge as has traditionally been supposed, this is a large difference between us and the bacterium indeed.[25] In LTOBC, I argued that negation, hence explicit contradiction, is dependent upon subject-predicate, that is, propositional, structure and vice versa. Thus, representations that are simpler also do not have propositional content.

In sum, these six differences between our representations and those of the bacterium, or Fodor's paramecia, ought to be enough amply to secure our superiority, to make us feel comfortably more endowed with mind.

NOTES

1. "Banish Discontent," in *Language, Mind and Logic*, Jeremy Butterfield, ed., (New York: Cambridge, 1986), pp. 1–23; J. A. Fodor, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press, 1987).
2. "Toward a Causal Theory of Representation," in *Contemporary Perspectives in the Philosophy of Language*, Peter French, Theodore Uehling Jr., and Howard Wettstein, eds., (Minneapolis: University of Minnesota Press , 1979), pp. 81–102.
3. "Misrepresentation," in *Belief: Form, Content, and Function*, Radu Bogdan, ed., (New York: Oxford University Press, 1986), pp. 17–36.
4. "Biological Functions and Perceptual Content," *The Journal of Philosophy* LXXXV, no. 1 (January 1988): pp. 5–27.
5. Millikan, Ruth. *Language, Thought and Other Biological Categories* (Cambridge: MIT Press, 1984) (hereafter LTOBC).
6. *Philosophy of Science* LVI, no. 2 (June 1989): pp. 288–302.
7. An odd custom exists of identifying this sort of view with Larry Wright, who does not hold it. See my "in Defense of Proper Functions." Natural selection is not the only source of proper functions. See LTOBC, chapters 1 and 2.
8. See LTOBC; and "Truth Rules, Hoverflies, and the Kripke-Wittgenstein Paradox," *The Philosophical Review* 99 (1990): pp. 323–53.
9. This last clarification is offered to aid Fodor ("On There Not Being an Evolutionary Theory of Content" [hereafter NETC], unpublished), who uses my term 'Normal' (here I am not capitalizing it but the idea has not changed) in a multiply confused way, making a parody of my views on representation. In this connection, see also fns. 13 and 17.
10. 'Normal explanation' and 'normal condition for performance of a function,' along with 'proper function,' are defined with considerable detail in LTOBC. The reader may wish, in particular, to consult the discussion of normal explanations for performance of 'adapted and derived proper functions' in chapter 2, for these functions cover the functions of states of the nervous system which result in part from learning, such as states of human belief and desire.
11. *Knowledge and the Flow of Information* (Cambridge: MIT Press, 1981).
12. Strictly, this normal condition must derive from a 'most proximate normal explanation' of the consumer's proper functioning. See LTOBC, chapter 6, where a more precise account of what I am here calling 'representations' is given under the heading 'intentional icons.'
13. In this particular case, one task is, surely, contributing, in conformity with certain general principles or rules, to practical inference processes, hence to the fulfillment of current desires. So, if you like, all beliefs have the *same* proper function. Or, since the rules or principles that govern practical inference dictate that a belief's 'shape' determines what other inner representations it may properly be combined with to form what products, we could say that each belief has a *different* range of proper functions. Take your pick. Cf. Fodor, "Information and Representation," in *Information, Language, and Cognition*, Philip Hanson, ed., (Vancouver: British Columbia University Press, 1989); and NETC.
14. These examples are of representations that are not 'inner' but out in the open. As in the case of inner representations, however, they are produced and consumed by mechanisms designed to cooperate with one another; each such representation stands intermediate between two parts of a single biological system.
15. Dretske (in "Misrepresentation," p. 28) and David Papineau (in *Reality and Representation* [New York: Blackwell, 1987], p. 67 ff) have similar concerns.
16. This is a broad metaphor. I am not advocating computationalism.
17. A word of caution. The normal conditions for a desire's fulfillment are not necessarily fulfillable conditions. In general, normal conditions for fulfillment of a function are not quite the same as conditions which, when you add them and stir, always effect proper function, because they may well be impossible conditions. For example, Fodor, in "Information and Representation" and NETC, has questioned me about the normal conditions under which his desire that it should rain tomorrow will perform its proper function of *getting* it to rain. Now, the biologically normal way for such a desire to be fulfilled is exactly the same as for any other desire: one has or acquires true beliefs about how to effect the fulfillment of the desire and acts on them. Biologically normal conditions for fulfillment of the desire for rain thus include the condition that one has true beliefs about how to make it rain. Clearly this is an example in which the biological norm fails to accord with the statistical norm: most desires about the weather are fulfilled, if at all, by biological accident. It may even be that the laws of nature, coupled with my situation, prohibit my having any true beliefs about how to make it rain; the needed general condition cannot be realized in the particular case. Similarly, normal conditions for proper function of beliefs in impossible things are, of course, impossible conditions: these beliefs are such that they cannot correspond, in accordance with the rules of mentalese, to conditions in the world.
18. *Psychosemantics* and NETC.
19. *Brainstorms* (Montgomery, VT: Bradford Books, 1978).
20. "Why Paramecia Don't Have Mental Representations," in *Midwest Studies in Philosophy* P. French, T. Uehling Jr., and H. Wettstein, eds., vol. x (Minneapolis: Minnesota University Press, 1986), pp. 3–23.
21. Accordingly, in LTOBC I did not call these primitive forms 'representations' but 'intentional signals' and, for items like bee dances, 'intentional icons,' reserving the term 'representation' for those icons, the representational values of which must be identified if their consumers are to function properly—see section V.5.
22. *The Philosophical Review*, XLV, no. 1 (1986): pp. 47–80.
23. Possibly human intentions are in both indicative and imperative mood, however, functioning simultaneously to represent settled facts about one's future and to direct one's action.
24. On the other hand, the bees cannot go two places at once.
25. In LTOBC, I defend the position that the law of noncontradiction plays a crucial role in allowing us to develop new methods of mapping the world with representations.

Inferentialism and Some of Its Challenges

Robert Brandom

II. What Is Inferentialism?

Logical empiricism revived classical empiricism by appealing to the new quantificational predicate logic Russell had developed out of Frege. Its motivating thought was that this way of understanding the inferential articulation of the immediate deliverances of sense provided powerful new expressive tools to put in the place of traditional appeals to processes such as association and abstraction, which had proven themselves woefully inadequate to rendering the contents of interesting empirical concepts—never mind mathematical ones. It is a striking fact about the contemporary scene that two broad classes of theories of concepts (I'll discuss a third in a moment) correspond to these two dimensions—sensuous and logical—into which the logical empiricists sought to factor conceptual content.

One popular strategy looks to the *observational* use of concepts as the key to conceptual content. Here one thinks of the use of 'red' or 'square' as non-inferentially elicited as a response, typically, to red or square things. The focus is accordingly on the reliable differential responsive dispositions linking, say, tokenings of 'horse' to horses. Fodor's and Dretske's semantic theories are principal examples of this class of approach. These deserve to be seen as contemporary descendants of classical *empiricist* theories of content. Another strategy is to look to the contents of *logical* concepts as providing the key to understanding conceptual content generally. Here the idea is to generalize Gentzen-style specifications of the meanings of logical connectives by pairs of introduction and elimination rules to notions of the circumstances and consequences of application of an expression. Dummett is the principal figure in this tradition, in which he is followed by others such as Peacocke, and myself in *Making It Explicit*. These deserve to be seen as offering specifically *logical* versions of traditional *rationalist* theories of content. Each is a self-consciously one-sided approach, by contrast to the even-handed appeal to both observation and logical inference in virtue of which Carnap's neo-Kantian roots become evident.

The genus of which semantic inferentialism is a species then has these features. It is:

a) categorially sententialist or top-down,
b) expressive, or sense-based, and
c) rationalist in its choice of conceptual paradigm,

in contrast to theories that are either categorially bottom-up, representational or reference-based, and empiricist in their choice of conceptual paradigm. Any theory of this genus faces three structural demands. Along the categorial dimension, it must show how to assign contents to subsentential elements. Along the semantic dimension, it must show how to underwrite a notion of reference or representation. And it must show how to model conceptual content generally, including especially the content of observational concepts, on that of the logical concepts it treats as paradigmatic.

What distinguishes inferentialist semantics within this genus is the concept it proposes to explore as a candidate for the notion of sentential sense treated as central by accounts of that genus and motivated by the case of logical concepts: inferential role. The idea is to understand propositional contents as what can both serve as and stand in need of *reasons*, where the notion of a reason is understood in terms of inference. So propositional contentfulness is taken to be a matter of being able to play the role both of premise and of conclusion in inferences. Once the notion of introduction and elimination rules as exhaustively constitutive of the content of logical concepts has been generalized to take in the circumstances and consequences of application of non-logical concepts, the step to inferentialism is taken when one understands their content as exhaustively constituted by the material, non-logical *inferential* connection between those circumstances and consequences. The content of a concept such as temperature is, on this view, captured by the constellation of inferential commitments one undertakes in applying it: commitment, namely, to the propriety of all

From *Philosophy and Phenomenological Research*, 2007, LXXIV: 3, pp. 651–76.

the inferences from any of its circumstances of appropriate application to any of its appropriate consequences of application.

Inferentialism is not the only way to try to develop an account that takes the contents of logical concepts as paradigmatic. (Peacocke, for instance, takes a somewhat different tack.[1]) But one might well ask what motivation there is for adopting this paradigm at all. Granted (though there is, to be sure, no unanimity even on this point) that Gentzen-style definitions offer us impressively clear and demonstrably useful specifications of the meanings of logical connectives, why should we think that that model can helpfully be generalized from logical to non-logical, paradigmatically empirical concepts? The drunk's reason for looking for his lost keys under the streetlamp—that the light is better there—notoriously does not provide a reason for thinking the keys are likely to be found there. Why doesn't a corresponding criticism apply to contemporary logical-rationalist approaches in semantics?

One reason is provided by looking at the motivations for a third contemporary candidate for a privileged subclass of conceptual contents on which to model the rest: *modal* concepts. For logical empiricism, like its classical Early Modern ancestors, also signally failed to render the contents of ordinary concepts. One retrospectively obvious reason is its lack of expressive resources sufficient to render the distinction between accidental generalizations and lawlike regularities—the very incapacity Kant had diagnosed already as fatal to the classical version of empiricism. The advent of a technically adequate semantics for modal concepts put philosophers in a position to create a third wave of empiricism. It had many of the same basic motivations and aspirations as the logical empiricism of the middle third of the twentieth century, but could use more powerful modal logical apparatus in place of the extensional logic previously appealed to as providing the logical cement binding together and articulating the sensuous content provided by perception. David Lewis may be taken as an index figure of this movement.[2]

What I want to focus on for the moment is the appreciation of the significance of what might be called 'non-Tractarian' concepts—paradigmatically modal, probabilistic, and normative ones—for our understanding of conceptual content generally. In an autobiographical sketch, Sellars says the central idea motivating his work was one that occurred to him already in the 1930s: that the centrality of modal concepts in formulating empirical ones—the insight he put into the lucid title of one of his most impenetrable essays "Concepts as Involving Laws, and Inconceivable Without Them"—meant that

> What was needed was a functional theory of concepts which would make their role in reasoning, rather than their supposed origin in experience, their primary feature.[3]

The expressive role distinctive of modal vocabulary is to make explicit the distinction between counterfactually robust inferences and those that are not—a distinction without which, Sellars reminds us, following Kant, we must fail to understand not only the content of such theoretical empirical concepts as rigidity or mass, but also such observational ones as red or horse. It is not at all clear how these modal features of empirical concept use might be understood in terms of chains of reliably covarying events linking horse-stimuli to 'horse'-responses. In this connection it might be noticed that although the liberal use of modal vocabulary in the metalanguage in which they conduct their explanatory enterprise is of the very essence of Dretske's and Fodor's semantics for the observational concepts they treat as paradigmatic of concepts in general, neither one so much as attempts to reconstruct within their theory the contents of those modal concepts. And the recalcitrance to empiricist reduction of what is expressed by modal vocabulary applies, *ceteris paribus*, equally to probabilistic and normative vocabulary. It provides the motivation for Sellars' inferentialism.

III. Varieties of Inferentialism

Inferential approaches to what is expressed by sentences can take various forms. What may be called '*weak*' inferentialism claims only that the inferential connections among sentences are *necessary* for them to have the content that they do, in the sense that unless at least some of those inferential involvements were as they are, the sentence would mean something different. By 'inference' in such a formulation is meant *correct* inference: the ones people *ought* to make, rather than the ones they are actually disposed to make, if those two notions fall apart. The claim of weak inferentialism, so understood, ought not, I think, to be controversial. For even those who understand what is said by sentences in terms ultimately of what is represented by

their subsentential components take it that their representational content *determines* the proprieties of inferences they are involved in. So at least some of the inferences are such that if they were not correct, the sentence would mean something else. For only if they had different truth conditions—for instance—could the sentences be understood as playing different roles as premises and conclusions of good inferences.

At the other end of the spectrum is what might be called '*hyper*inferentialism.' It is the claim that the inferential connections among sentences, *narrowly construed*, are *sufficient* to determine the contents they express. By 'narrowly construed' in this formulation is meant that only correct inferences in which the sentence plays the role of premise or conclusion are to be considered. Hyperinferentialism is—Gentzen claims, and I think we ought to agree—true for *logical* vocabulary. The introduction rules for logical connectives appeal only to inferential grounds for applying them (as principal connectives) that are sentences in which the connective being defined or introduced does not occur (as principal connectives). And their elimination rules appeal only to inferential consequences of applying them (as principal connective) in which the connective in question does not occur (as principal connective). But hyperinferentialism is extremely implausible as applied to other sorts of vocabulary, especially to vocabulary that has *observational* uses that are essential to its meaning. The meaning of 'red' is not just a matter of what other concepts its applicability is properly inferrable from (e.g., scarlet), and the applicability of what other concepts is properly inferrable from its applicability (e.g., colored), nor of the applicability of what other concepts preclude its applicability (e.g., green or multiple of 3). Its *non*inferential applicability to *red things* is also an essential propriety of the use of the concept red: one that must accordingly be underwritten by any adequate account of the meaning or content expressed by the use of the word 'red.' If taking the meanings expressed by logical vocabulary as a model for the semantics of other sorts of expressions could take the form only of commitment to hyperinferentialism, then there would be little reason to take this explanatory strategy seriously.

But there is a *via media* between the uncontroversial but unhelpful truism of weak semantic inferentialism and the powerful and interesting but unsustainable (outside the realm of logical and perhaps some mathematical discourse) thesis of semantic hyperinferentialism. What may be called '*strong*' inferentialism claims that the inferential articulation of concepts, *broadly* construed, is *sufficient* to determine their contents. By 'broadly construed' in this formulation is meant three things. First, the inferences in question must be understood to extend beyond *logically* or *formally* good ones—those whose correctness is settled just by the logical form of the sentences involved. They must include also those that are *materially* correct—that is, those that intuitively articulate the contents of the *non*logical concepts involved. Sellars offers as examples the inference from 'A is to the East of B,' to 'B is to the West of A,' and from 'Lightning is seen now,' to 'Thunder will be heard soon.' Second, besides material inferential relations among sentences in the sense of their proper role as premises and conclusions, material *incompatibilities* among sentences, which underwrite inferential relations in a narrower sense, are included. Thus the fact that the correct applicability of square precludes the correct applicability of triangular, so that the inference from square to *not*-triangular is a good one, is also to be considered.

Third, and most important for understanding the difference between the hyperinferentialist and strong inferentialist theses in semantics, is that inferential relations between *non*inferential circumstances of appropriate application and *non*inferential appropriate consequences of application are also taken into account. The way the Gentzen hyperinferentialist model for the semantics of logical concepts is to be extended is by taking seriously the thought that in using *any* expression, applying *any* concept, one is undertaking a commitment to the correctness of the (in general, material) inference from the *circumstances* in which it is correctly applied to the correct *consequences* of such application. And this is so even where some of those circumstances or consequences of application are *non*inferential. Thus the visible presence of red things warrants the applicability of the concept red—not as the conclusion of an *inference*, but *observationally*. And the point is that the *connection* between those circumstances of application and whatever consequences of application the concept may have can be understood to be *inferential* in a broad sense, even when the items connected are not themselves sentential. In a culture in which **white** is the color of death, and things associated with death are to be shunned or avoided—a culture, to be sure, that would mean something somewhat different than we do by their word corresponding to our 'white'—the connection

between the visible presence of white things and the practical response of shunning or avoiding, which their practitioners endorse by using the concept in question, is an inferential one in the broad sense in question here.

It is strong semantic inferentialism that is articulated and endorsed in *Making It Explicit*. The two key moves in extending the inferential approach beyond its paradigmatic application to logical concepts are first, looking to *material* inferential and incompatibility relations, and second taking into account the *inferential* relations linking circumstances and consequences of application, even where these are *non*inferential circumstances or consequences.[4] The generic location of this approach in the botanization presented in the first two sections of this paper—as categorially top-down, and semantically expressive rather than representational—then dictate the principal constructive obligations of inferentialist approaches in semantics. It must be extendable somehow to *subsentential* expressions, such as singular terms and predicates, quantifiers and so on, which cannot play the directly inferential roles of premise or conclusion in inferences. And it must somehow underwrite assessments of the *representational* content of expressions of all categories: the truth of sentences and the reference or denotation of terms, predicates, and so on. In *Making It Explicit*, the first of these tasks is addressed by using Frege's methodology of marking *inferential* invariances under *substitution* of one expression for another. Roughly, two subsentential expressions play the same *indirectly* inferential role just in case substituting one for the other never turns a good inference (in the broad sense) into a bad one. The second task, offering an account in inferential terms of the representational dimension of content, is addressed (in Chapter Eight of *Making It Explicit* and Chapter Five of *Articulating Reasons*) by appealing to the different *social perspectives* corresponding to the difference between the *practical deontic attitudes* of *undertaking* a commitment (oneself), and *attributing* it (to another), which are made explicit in the difference between *de re* and *de dicto* ascriptions of propositional attitude.

This constructive task of semantic theories of inferentialist shape—roughly, underwriting representational semantic characterizations in terms of inferential ones, or understanding reference in terms of inference—points to a further division. For another important way in which it is useful to characterize varieties of semantic inferentialist programs concerns the nature of their methodological aspirations. The big distinction in this vicinity is that between *reductive* versions of inferentialism and *expressive* ones. As I would understand it, reductive inferentialism would claim first that there can *be* expressions or intentional states standing in inferential relations and playing inferential roles, hence having conceptual content, without yet standing in representational ones, and second, that one can then *build* representational relations and roles, and so content, out of those inferential ones. Compare, for the case of the converse order of explanation, a story where representational relations are defined in terms of nomological relations between representings and representeds, and a story is told only *much* later about how they must interact to yield representations of states of affairs that are truth-evaluable (hence believable), and so inferentially related.

Expressive inferentialism, by contrast, is a claim about *understanding* inferential and representational relations. It is at the level of the *senses* of the concepts inference and representation, rather than at the level of their *referents*. The expressive inferentialist acknowledges that nothing can stand in genuinely *inferential* relations unless it also has *representational* content. There is no inference without reference. But it is claimed nonetheless that one can specify sufficient conditions for expressions to be used so as to possess conceptual content (of *both* sorts) in a purely *inferential* metalanguage. Seeing what it is about this inferential articulation that amounts to the possession of representational content is then explaining what is expressed by the representational semantic meta-language (which turns out always already to have been in principle applicable) in the terms of the inferential semantic metalanguage. So one of the most important concluding moves developing the expressive inferentialism in *Making It Explicit* (reprised in Chapter 5 of *Articulating Reasons*) is an account of the broadly inferential role locutions must play in order for them to mean what 'of' and 'about' do when expressing the representational dimension of intentionality—the uses of 'of' and 'about' epitomized by 'thinking of a horse' and 'talking about colors,' rather than 'the pen of my aunt,' and 'weighing about five pounds.'

IV. Which Inferences?

It is obvious that a key question for strong expressive inferentialism—as for any species of this genus of semantic explanatory strategy—is

(as Fodor and Lepore put it in a recent article[5]) 'Which inferences are meaning-constitutive?' I want to discuss briefly two kinds of response to this question (without meaning to deny that other avenues are also open). A natural way into the issue is provided by one of Quine's central challenges in "Two Dogmas of Empiricism." Transposed from the idiom of analytic *truths*—sentences whose truth is underwritten by the meanings of the nonlogical expressions occurring in them—to one of, as it were, analytic *inferences*, i.e., those whose correctness is underwritten by the meanings of the nonlogical expressions occurring in them, the challenge is this. Semantic theories associate contents (or other semantic interpretants) with expressions, performances, or states in order to explain or at least codify proprieties of the *use* of those expressions, the practical *significance* of those performances, or the proper *functioning* of those states. Specifications of the distinction between correct and incorrect uses accordingly stand to attributions of content very much as statements in an observation language stand to statements in a theoretical language in ordinary empirical science: the *point* of the theory is to explain what is observable, and the point of semantics is to explain practice. So if one of the principal moves proposed by a semantic theory is to distinguish two flavors of inference (or truth), namely those that are constitutive or expressive of meaning, and those that are good rather in virtue of something else, we are entitled to ask what feature of the *use* of expressions with those meanings or contents reflects this semantic difference. (One need not be committed to semantic instrumentalism to ask this question, any more than one need be committed to instrumentalism more generally in order to ask what observable phenomena are explained by or manifest an hypothesized, unobservable, theoretical state of affairs.)

Quine, of course, surveys a number of candidates such as unrevisability, and finds all of them wanting. He famously concludes that postulating meanings, and thereby committing oneself to some inferences (truths) having a semantic privilege that not all do, can do no explanatory work. Sellars offers a direct response to this challenge. The practical difference that Quine rightly demands is, according to Sellars, just the difference between counterfactually robust inferences and those that are not. The inference from something's being copper to its melting at 1083.4°C is partly constitutive of the concept copper because *if* the coin in my pocket *were* copper, it *would* melt at that temperature. The inference from the coin's being in my pocket to its being copper is *not* partly constitutive of the concept copper because *if* this nickel *were* in my pocket, it would *not* follow that it was copper. This is a directly responsive answer to Quine's challenge, because we do in our ordinary linguistic practice distinguish between inferences based on their modal status as counterfactual-supporting or not—between those that would explicitly be licensed by law-like regularities such as connect atomic structure and melting point, and those that would be licensed only by accidental regularities, such as those connecting atomic structure and location in space. For Sellars, conceptual connections are just the lawful ones. (That is why 'concepts involve laws and are inconceivable without them.') His was the first explicitly *modal* theory of meaning.

To be sure, this approach has some radical consequences. If we are wrong about the laws of nature, then not only have we gotten the facts wrong, we are using incorrect *concepts*. Investigating the world is an attempt at once to rectify our claims and our concepts. Conceptual change is part and parcel of scientific change, because every new law we discover and every old one we are forced to give up brings with it a change in our concepts. Semantics is not a discipline that can be pursued independently of our empirical inquiries into the rest of the world. These consequences may not sit well with our pretheoretic talk about meaning, but so what? Our ordinary talk of hot and cold runs together differences in temperature and differences in specific heat ('Stone floors are colder than wood floors,'[6]) but we wouldn't want to hold our theoretical physics hostage to our casual, unreflective practice. Why do so in semantics?

Another approach is Quine's own: put *all* the inferences in the same basket, by rejecting the distinction between those good in virtue of meaning and those good in virtue of matters of fact. Such an approach can still accommodate a notion of sentences as contentful—as, we can say if we like, expressing contents. For sentences can still be understood as playing roles as premises and conclusions of inferences. But since for the most part the inferences in question are *multipremise* inferences, there is a sense in which no sentence plays its inferential role all on its own. For what a claim is evidence for or against and what is evidence for or against it depends on what collateral commitments are available to serve as auxiliary hypotheses in

extracting inferential consequences. The inferential significance of endorsing any one sentence depends on what else one is committed to. This is the line of thought that leads Quine to think that 'the unit of meaning is the whole theory,' rather than individual sentences. This holist consequence goes against the grain of the semantic atomism that has been at the center of the empiricist tradition—in both its traditional and its distinctive twentieth century logical and modal forms. But that incompatibility doesn't by itself show that the holist response is incoherent.

I think that both these strategies are open to contemporary semantic inferentialists. Sellars's approach seems to be wholly viable, though it has not, as far as I know, yet been pursued by other theorists. In *Making It Explicit*, I adopt a version of the Quinean strategy. This does in a certain sense involve giving up the notion of content—at least, it means giving it up in any sense that would be recognizable to a semantic atomist. But the suggestion is that we can do without talk of contents and meanings as *things* associated with sentences one by one in favor of talk of inferential connections among sentences that are contentful precisely in virtue of standing in those relations. The question of whether 'all of them' is, like Sellars's different one, a responsive and viable answer to the question: 'Which inferences matter for content?' then becomes the question of whether one can make sense of semantic holism.[7] So let us look at that issue.

V. Holism

The biggest challenge for holistic semantic theories has always been accounting for the possibility of *communication* or of interpersonal *understanding*. If the inferential significance of a claim depends on what else one is committed to, then any difference between the collateral commitments of speaker and audience can mean that a remark has a different significance in the one's mouth than it does in the other's ear. How is it then possible to make sense of the idea that they understand one another, so as to be able to agree or disagree? If the contents expressed by sentences must be individuated as finely as the theories they are embedded in, the intelligibility of communication across theories—the very notion of conveying information—is threatened. And the issue arises as urgently for diachronic communication as for synchronic, face-to-face cases. If, because of his very different collateral commitments, Rutherford meant something quite different by 'electron' than I do, it seems I can't disagree with him about whether electrons have fixed positions and orbits, since I can't either say or think anything with the content he would have expressed by saying 'electrons orbit the nucleus.'[8] How, then, are we to understand so much as the possibility of cognitive progress in science?

Quine, of course, fully appreciated the force of these challenges. It is precisely such considerations of semantic incommensurability of meanings that led him to insist that concepts such as communication, information, and cognitive progress—most importantly as they matter for the serious business of the conduct of science, but also in our more informal transactions with each other—must be understood as arising at a different semantic level: that of *reference*, rather than *meaning*. Our starting-point in this essay was the commonsensical observation that besides what is said or thought there is also what is talked or thought *about*. Though Rutherford and I may be saying different things—expressing different contents—when we use the phrase 'the location of the electron,' we can both be talking about or referring to the same things: electrons. And we may be classifying them as falling within the same extension when we characterize them as 'fast-moving.' Given his collateral beliefs concerning its divinity, the Zoroastrian sun worshipper surely means something different by the term 'the Sun' than I do, but we may still both believe *of* the Sun and *of* visible things that the first is to be counted among the second. Or, if we don't already share it, he can convey that information to me, in spite of our other semantic differences.

Quine's way out, then, is to neutralize the otherwise corrosive effects of holism at the level of what is said or thought by appeal to the representational dimension of what is talked or thought about. Coreference or coextensionality is an equivalence relation that maps the disparate meanings expressed by sentences animated by distinct theories onto one another in just the way needed to underwrite the possibility of communication or information, reciprocal understanding, and so on. Again as noted at the outset of this essay, any semantic theory that begins with 'that' intentionality must eventually explain its relation to 'of' intentionality in any case—must proceed from an account of *senses expressed* to one of *objects* (and sets of

objects) *denoted*. So it is open to any theory that adopts this order of semantic explanation to adopt Quine's strategy of appealing to what is talked or thought about to secure an account of the nature of communication.

One might, it is true, at this point ask what work the holist theory of meaning (the expressive dimension of intentionality) would be doing, if it is immediately abandoned in favor of the representational dimension of intentionality when hard questions are asked about communication and shared understanding. But the answer on which the adoption of the order of explanation being considered is premised is clear: it is at that level that we are to understand the use of expressions (the functioning of states or performances) in virtue of which we can understand them as having representational properties and relations. The account in *Making It Explicit* is of this sort. The explanation of what is expressed by declarative sentences, and so by the 'that' clauses of *de dicto* ascriptions of propositional attitudes, is offered in inferentialist terms. Then substitution-inferential commitments and their anaphoric inheritance are shown to be sufficient to explain what is made explicit in the *de re* portions of such ascriptions: the part where it is specified what one is speaking *of* or thinking *about*. Thus the implicit representational dimension that is expressed explicitly by the use of terms such as 'of' and 'about,' which is invoked when we say that both Rutherford and I were thinking of electrons, and that the Zoroastrian was talking about the Sun, is explained in ultimately inferential terms. Inferentialists do not, and should not want to, deny the existence or the importance of the representational dimension of intentionality. Rather, they are committed to an order of explanation that seeks to *understand* 'of' intentionality in terms of 'that' intentionality. Coreference—a kind of intersubstitutability of expressions—then provides a respect of similarity across inferential roles, grouping them into extensional equivalence classes, which interlocutors *can* share.

In his writings on holism, Fodor has argued that where meanings-as-inferential-roles are individuated so finely that there is little hope of different interlocutors (or even one interlocutor at different times) having *identical* ones, it is of no use for the holist semantic theorist to attempt to retreat to the idea of mutual understanding consisting in the grasp of at least *similar* meanings, or meanings that in some sense 'overlap.' One might initially think that, for example, Rutherford and I at least share *some* of the inferential consequences of application of our uses of 'electron.' We both agree, for instance, that it follows from something's being an electron that it is negatively charged, has a mass that is orders of magnitude smaller than that of the proton, that its movement creates a magnetic field in the direction specified by the right-hand rule, and so on. But that overlooks the fact that 'charge,' 'proton,' and 'magnetic field' all by hypothesis *also* mean something different in his mouth than in mine. Once we realize that these are all in exactly the same boat as 'electron,' we see that we've just put the issue off, rather than solving it by defining a sense of 'similarity' that consists in having *some*, but not all, inferential antecedents and consequences in common. We actually have *nothing* in common.

But this conclusion is too strong. For we do share the *words*, at least in the sense of noise- or sign-design types. When Rutherford sees lightning, he, like me, is committed to the correctness of applying 'electron'; and when he does apply it, he, like me, is committed to the correctness of the application of 'charge,' 'magnetic field,' and so on. One might respond at this point on Fodor's behalf that although *some* kind of similarity metric is induced by counting the *noises* that express the conclusions two interlocutors would draw from, or the promissory claims they would count as evidence for, claims expressed using 'electron,' still that is only because we have restricted ourselves exclusively to *nonsemantic* properties of their utterances. So nothing like shared *meaning* is thereby underwritten. But once again, this is too hasty a conclusion. Here one might think of Davidson's account of communication as interpretation. Davidsonian interpretation is explicitly understood as consisting in mapping the *noises* made by the interpretive target onto the noises made by the interpreter. In understanding another, I am to use sentences in *my* mouth to attach as labels to his sentences, and thereby serving to 'measure' them. My interpretation is a useful one—I have understood what the other says—just insofar as the inferential moves I am committed to endorse with the noises I produce mirror (perhaps with a qualifying commentary couched in my own idiom) his. The ideal interpretation is a homomorphism, a structure-preserving mapping, from his noises onto mine, preserving the consequential structures. Where not all the moves between his noises that he is committed to the correctness of are matched by similar commitments on my part regarding

my counterpart (according to the interpretation mapping) noises, only similarity, and not identity of consequential role—and so only partial understanding—is achieved. And recall that inferential role in the broad sense includes the inferential connections between circumstances of appropriate application and appropriate consequences of application quite generally—even when the circumstances or consequences of application themselves are *non*inferential. Thus Rutherford and I are both disposed to respond to a bolt of lightning by applying the term 'electron,' and to respond to applying the expression 'high voltage, high amperage electron flow' to a bare piece of metal by avoiding contact with it. These language entry and language exit moves, no less than the language-language ones, also give us something important in common, even when described at a so-far-subsemantic level, that is, in a nonsemantic vocabulary. I do not see why the structures so-described do not underwrite a perfectly intelligible notion of *partially* shared, or merely *similar* inferential roles.[9]

A more radical response than Quine's, one that obviates the necessity of considering similarities rather than identities of meaning as what is shared, is one that rejects the model of sharing on which the worry about the intelligibility of communication in the face of the holistic character of meaning is based. The motivating picture is at root a Lockean one: the speaker has an idea in his head, and his uttering the words he does succeed in communicating that idea to me if the idea hearing them uttered arouses in me is the same (repeatable) as the one he has. The meaning or content is, as it were, to be transported from his head to mind, or reproduced in mine. But one could think of understanding rather on the model of a cooperative practice or activity. In particular, I can be said to understand your remark insofar as I can compute its inferential significance both for you and for me, and navigate successfully back and forth across the two perspectives on its content constituted by the background of auxiliary hypotheses drawn from *your* collateral commitments and the ones drawn from *mine*.

To do that, I need to be able to determine what you would take to be the consequences of your claim, and what would be evidence for it and against it, given *your* theory of things. This will matter if I want to predict what else you'll go on to say or believe, or what you will try to do in a particular situation. (If you believe that animal over there is a deer and desire to shoot a deer, then you may try to shoot the animal.) I also need to be able to determine how *I* should draw inferences, using *your* utterance as a premise. That includes mapping your noises onto mine, and then extracting the inferential consequences from the claim that would be expressed by *my* corresponding sentence utterance, with the help of auxiliary hypotheses drawn from other claims *I* endorse. This will matter if I want to extract *information* from your remark, that is, to draw conclusions from it about what is *true*, and if I want to predict what you are likely to *succeed* in doing. (If you *do* try to shoot the animal, I know, though you do not, that what you are likely to succeed in shooting is in fact a horse.)

The capacity to understand each other is the practical ability to navigate across the gulf between doxastic perspectives created by the effect of differing collateral commitments on the inferential significance of one noise in the mouth that utters it and the ear that hears it. When that implicit skill is made explicit in the form of claims (thinkables, believables, sayables), it takes the form of *de dicto* and *de re* ascriptions of propositional attitudes. 'The speaker believes *that* that animal is a deer,' 'The speaker believes *of* or *about* a horse that it is a deer.' The expressive role of explicitly representational terms such as 'of' or 'about' in this usage is to mark that the identity claim (intersubstitutional commitment) connecting 'that animal' and 'a horse' is one that the ascriber *undertakes*, and is using to express the content of the propositional commitment *attributed* to the speaker. When what is at issue is what *I*, the ascriber, should conclude from your remarks, I am going to map your 'deer' talk onto my 'horse' talk. I *say* that (rather than just *doing* it), precisely by using *de re* forms of ascription: talking about what you are talking *about*, what you are speaking *of*, what you *represent* as what. Those *de re* ascriptions are the home language games of such explicitly representational locutions—the ones that make explicit the representational dimension of intentionality, the 'of' flavor, rather than the *de dicto* 'that' flavor. That is, it is their use in such contexts that *makes* them explicitly representational locutions. And that expressive role can be understood in terms of the *inferentially* articulated and specified distinction of perspective between premises attributed and those endorsed. In that way adopting the navigation-across-perspectives model of communication can also be seen as a way of developing the Quinean retreat-to-reference response to the challenge to make communication

intelligible within a holistic semantics of what sentences express.

Practitioners who understand each other in this practical sense—who can successfully make use of each other's remarks in their own reasoning, both about what the other has reason to do (given his or her beliefs) and about what one has oneself reason to do (given one's own beliefs)—do indeed 'share' something. But what they share is like the dance that Fred and Ginger are doing together—one and the same dance, even though individually they are doing different things (him going forward, her backward; she dipping, he holding; she twirling, he leaping). It is not like the cadence that the soldiers marching in step share: something visible already in what *each* is doing individually, simply repeated across them all. We can think of conceptual (that is, inferentially articulated) content as like Fred and Ginger's dance: something that is *essentially perspectival*, in that grasping it (like engaging in the dance) requires doing *different* things from each individual participating (in the conceptual case, depending on the background constituted by their other commitments). This is a different model of understanding and communication from the Lockean repetition or reproduction model—the soldiers marching model. An account of this shape will count as leaving communication unintelligible only to those who insist upon the Lockean model as the *only* way to understand understanding each other.

Besides the Quinean retreat-to-reference response and the practical-navigation-across-perspectives response, there is a third way in which inferentialism can (and the theory expounded in *Making It Explicit* does) respond to worries about the effects on the intelligibility of communication of the relativity of inferential significance to collateral commitments. For another great division among theories of concepts and their contents is that between broadly Cartesian and broadly Kantian approaches. Cartesians think of concepts as something like mental particulars. The principal question about them concerns the thinker's grasp of them: how well do we really understand them. Kantians think of concepts rather as norms or rules that bind those who apply them, determining what would count as a *correct* judgement in which they are applied, or a *successful* intention in which they were applied. For them the principal question does not concern our grip on the concepts, but rather their grip on us. So for Kant, issues of the *bindingness* of concepts, the way in which they become *valid* (gültig) for a thinker and agent, is the central philosophical issue. Transposed into a linguistic key, the question becomes what I must do in order to count as having applied, say, the concept copper, in thought or assertion—to have subjected myself to assessments of the truth of my claim or the success of my action accordingly as what I am talking about or acting on is or is not *copper* (rather than, say, just some reddish metal).

A good model of the second kind is playing a counter in a game. Once I count as a player in the game, I can play a counter that *has* a certain significance—obliges me to make some further moves under various circumstances, precludes me from making others, entitles but does not oblige me to make others, and so on. And the facts about the normative significance of that move may significantly outrun what I *understand* that significance to be. I may not realize all, or even very many, of the aspects of the normative significance of my performance, for it nonetheless to have that significance. I do not, for instance, need to *know* that the melting point of copper is 1083.4°C in order to call something copper, and thereby to have committed myself to its not melting at 1083°, but melting at 1084°C—in the sense of having said something that is true if and only if that condition obtains. Thus my remark or thought is subject to assessment according to that norm, even though I may not be aware of that fact.

On a Kantian picture, then, you and I can share a concept even though our *dispositions* to call something 'copper' are quite different—perhaps because of our different collateral commitments. In spite of such differences, we may be understood as binding ourselves by the very same complex norm for assessments of truth and success by our use of the word 'copper.' For what matters for such assessments is what auxiliary hypotheses (e.g., about the melting point of copper) are *true*. The fact that you and I have different views about which these are, and so are disposed to draw different conclusions from something's being copper, does not alter what *really* follows from it. *De dicto* specifications of the content of another's thought depend on the inferences she is disposed to draw from it: what she *thinks* she is committing and entitling herself to thereby. They articulate her *conception. De re* specifications of the content of another's thought, by contrast, depend on what inferences to and from it are *in fact* (including counterfactually) correct. They articulate the *concept* she has applied, even insofar as she is

ignorant of or mistaken about its content. The practical navigational capacities that are made explicit in *de re* specifications of the contents of ascribed propositional commitments express the standing commitment each of us has to their being *one* set of inferential roles that bind *all* interlocutors: those, namely, determined by multipremise inferences in which the collateral commitments supplying auxiliary hypotheses are *true*.

I conclude that the sensitivity and relativity of the inferential significance of a sentence to collateral commitments poses a threat to the intelligibility of communication *only* for a theorist whose own collateral commitments at the metalevel include *all* of:

a) commitment to communication's having to take place at the level of meaning, rather than of reference;
b) commitment to a Lockean reproductive model of the sort of sharing that communication consists in; and
c) commitment to a Cartesian, rather than a Kantian model of our relation to concepts.

The inferentialism of *Making It Explicit* explicitly rejects all of these assumptions. (If you think the couch cannot possibly go on the wall next to the fireplace, that may be because you are not thinking about moving enough of the rest of the furniture. Almost certainly you are not thinking about knocking out one of the interior walls.)

VI. Compositionality

A related worry about inferentialism concerns the *productivity* of language and thought: the fact that competent speakers and thinkers are able to produce and understand an indefinite number of sentences that express *novel* contents—not just novel in the sense that that speaker, hearer, or thinker has never entertained them before, but in the much stronger sense that *no-one* has ever done so. This striking observation was first made by Chomsky almost fifty years ago, and it has been adequately confirmed since in many ways—ranging from statistical analyses of empirical corpora of actual utterances to theoretical analyses based on the sentences of given lengths generated by particular partial grammars given fixed lexicons. About the only idea anyone has as to how to explain this striking fact is to treat language as *compositional*, at least in the weak sense that semantic interpretations of unfamiliar sentences are understood as generated by operations on semantic interpretations of their familiar parts.

The compositionality challenge to inferentialism arises because it is essentially a categorially top-down order of semantic explanation. It begins, not with the contents of subsentential components, but with what is expressed by whole declarative sentences. The thought is that (as a recent paper puts it): 'productivity demands compositionality, and compositionality implies the priority of subsentential semantics to sentential semantics.'[10] The first of these claims ought to be granted (at least for a suitably broad understanding of 'compositionality'). But the second is surely too strong. Compositionality does not imply semantic atomism, but at most what Dummett calls 'molecularism.' A molecularist order of semantic explanation starts with sentences, and so contrasts both with fully holist theories, which start with whole idioms or theories, from above, on the one hand, and atomist theories, which start with subsentential expressions such as singular terms and predicates, from below. That is, it takes seriously the idea that the sentence is the minimal unit for which one can take *responsibility*, or which can express the undertaking of a *commitment* (Kant), the minimal unit to which pragmatic *force* can attach (Frege), and the minimal unit with which one can make a move in a language game (Wittgenstein). The notion of semantic content is accordingly introduced to begin with as a way of codifying or explaining proprieties of the use of expressions of this category: sentences. (One might or might not look to *inferential* properties of the sentences at this point in the story, as inferentialists do.) One might do this for a relatively small finite corpus of sentences—say, those which a given interlocutor has used either as speaker or hearer, during his initiation into the language. One then looks to codify or explain features of the use of those sentences by noting similarities among them, paradigmatically that they contain occurrences of the same subsentential expression. Looking at the contribution the occurrence of that subsentential expression makes to the features of the use of the whole sentences in virtue of which they are intelligible as semantically contentful then allows one to attribute semantic content, in a derivative, indirect sense, also to those subsentential expressions. That in turn can allow one to generate semantic interpretants for a much larger class of sentences, compounded in familiar ways out of the familiar

parts. Compare: offering a structural analysis of the behavior of ordinary solids, liquids, and gases in terms of their molecular composition, and only later seeking to explain the behavior of those molecules, and many others one has *not* observed, on the basis of their atomic structure.

Like Dummett, I think that Frege pursued such an explanatory strategy of *de*composition and *re*composition. He starts with sentences, as the bearers of truth values, and assigns Bedeutung to singular terms only so as to be able to analyze the sentences as applications of functions to arguments. His strategy for specifying the truth conditions (and inferential roles) of arbitrarily iterated quantifications depends on one being able to form complex predicates by substitutional variation of whatever sentences are already on board, and then forming new sentences from those predicates by clamping quantifiers onto them. This way of proceeding does not even depend on the original sentences having literal lexical-syntactic parts; it is enough if they can be sorted as similar to one another in respects that act enough like the equivalence classes generated by orthodox substitutional variants.[11] There may be various reasons not to want to adopt the method of decomposition and recomposition. But it is surely a coherent strategy for achieving compositionality. So achieving the effect of compositionality does not *imply* the explanatory priority of sentences to subsentential structures.

The sort of inferentialism pursued in *Making It Explicit* explains our capacity to understand novel sentences in two different ways. First, it includes an account of the introduction of *logical* vocabulary in a way that is straightforwardly and traditionally compositional. Logical vocabulary is demarcated by its playing a distinctive expressive role: making *explicit* in propositional (claimable, believable, thinkable) form broadly *inferential* commitments that otherwise remain implicit in practical assessments of practice. The paradigm of logical vocabulary in this sense is the conditional, which lets one *assert* an inferential connection between (what is expressed by) the sentences that appear as its antecedent and consequent. In the same way, negation can be understood as codifying commitment to the *incompatibility* of claims. But on this account, *normative* vocabulary also qualifies as logical, in that it serves to codify commitment to the correctness of various patterns of *practical* reasoning. And other bits of vocabulary, from identity and quantificational locutions to the 'of' and 'about' that express representational intentional directedness (like normative locutions in that they are not usually thought of as distinctively *logical* vocabulary) show up as playing expressive roles of these kinds. For each, it is shown how they can be introduced into languages that do not already contain them, in such a way that the inferential roles of sentences in which they appear are settled by the inferential roles of non-logical sentences formulated in the prior vocabulary. It is in these terms, for instance, that the inferential roles of arbitrarily iterated mixtures of *de dicto* and *de re* ascriptions (sentences such as 'Tom believes of Benjamin Franklin that Henry Adams believed that he was not the inventor of the lightning rod,') are computed. For the special class of logically compound sentences (including those formed by using *sub*sentential logical locutions), this account is compositional in the classical bottom-up sense, albeit conducted within an inferentialist framework.

The inferences (and so inferential roles) generated for compounds formed by applying vocabulary that is *logical* in this special, semantically expressive sense are not restricted to those that are good in virtue of logical *form* (even for classical logical particles like conditionals and negation), because of the special expressive role discerned between the use of *non*logical sentences and the logical compounds that make explicit aspects of that use. They are, however, all inferences that are good in virtue of the (broadly inferential) *meaning* or *content* expressed by the use of the logical vocabulary in question. But not all vocabulary is logical vocabulary even in this extended semantically expressive sense, and (so) not all material inferences manifest the contents of logical concepts. What about the rest?

Here the account in *Making It Explicit* adopts a version of the decomposition-recomposition strategy. The material inferential behavior of nonlogical sentences is dissected with a substitutional scalpel by using Frege's basic methodological idea that what it is for two subsentential expressions to play the same indirectly inferential role ('indirect' since they don't express reasons in the sense that they cannot play the directly inferential roles of premise and conclusion of inferences) is for substitution of one for the other to preserve (important features of) the inferential roles of the sentences they occur in. Equivalence classes of subsentential expressions generated by these relations can then be associated with what are called (in Chapter Six of *Making It Explicit* and Chapter Four of

Articulating Reasons) 'simple material substitution inferential commitments.' Those subsentential, indirectly inferential contents then determine the correctness of all the substitution inferences involving compound expressions in which they occur.[12] This is how compositionality works for material inferences: via substitutions. It can be shown that the formal power of this substitutional decomposition and recomposition is equivalent to that of the standard functional-categorial grammars that David Lewis considers in 'General Semantics.'[13] That is, the result is as compositional as the most powerful approaches that assign semantic interpretants to, say, singular terms and sentences (let us say, objects and sets of possible worlds), and then generate semantic interpretants for derived categories functionally: predicates as functions from (tuples of) objects to sets of possible worlds, adverbs (such as 'slowly') as functions from functions from objects to sets of possible worlds to functions from objects to sets of possible worlds, and so on.

That is not the end of the story, however. For once again, not *all* the material inferences sentences are involved in can be computed on the basis of substitutions or the functions they generate. There will be a significant residue of multipremise material inferences whose correctness is *not* settled in this way by the contribution made to the inference potential of the various individual components that occur in the premises and conclusion, even when they are all summed. New things happen. That the apple is red does not follow either from its being ripe or from its being a Macintosh, but does follow from both premises together. The inference that would be explicitly licensed by the conditional 'If an apple is *both* ripe and a Macintosh, then it is red,' is not a substitution inference. If that claim is true, then it does articulate part of the commitment that is implicit in applying the terms 'Macintosh,' 'ripe,' and 'red' and in that sense is a feature of the contents they express. But it is not a consequence of features of their use that are manifested one by one. In this sense (though, given the rejection of a distinction between inferences made good by meaning and inferences made good by how the world is in favor of seeing them all as having both sources—for very different reasons), I agree with Fodor and Lepore that 'in general, the inferential role of a sentence/thought is not determined by the inferential roles of its constituents.'[14] But I think this fact evidences not a particular defect of inferentialism, but simply a fact about languages and (so) concepts. It is important not to treat languages as more compositional than they are. They are compositional with respect to their substitution inferences, but not with respect to the rest.

Some material proprieties of multipremise inferences cannot be discovered simply by inspecting the use of their component expressions. The substitution inferences give us a good handle on the proprieties of use governing novel compound sentences. They ensure that we always have a place to start in sorting the inferences involving the novel sentences into good ones and bad ones. But since what inferences are good—and so, on this line, what our words mean and what content our thoughts have—depends on how the world actually is (for instance, on what color ripe Macintosh apples are), we may have to go out into the world to find out what follows from or is evidence for or against novel claims. The idea that this sort of failure of compositionality is a *problem* for semantic theories is a product of commitment to semantic theories having to be both categorially bottom-up, and a Cartesian transparency thesis about the epistemological availability of the contents we grasp in the sense of being able to deploy them in thought. But these commitments are optional, and are rejected by inferentialists.

VII. Conclusion

In the first three sections of this paper, I sought to place inferentialism in a more general botanization of semantic theories, and to distinguish various more specific shapes versions of that approach can take. In the second half of the paper, I considered inferentialist responses to worries about its capacity to distinguish the inferences constitutive of conceptual content, about the intelligibility of communication and reciprocal understanding given its attendant holism, and about its ability to deal with the phenomena of productivity and (so) compositionality. Those interested in more detailed workings-out of those responses will find some in *Making It Explicit*. But I don't take it that the arguments on offer there settle the ultimate viability of the specific version of inferentialism they articulate. I've told the story in the way I have here in part as a reminder of how misleading it can be to assess the claims of a systematic theory piecemeal. Very often one move makes sense only in the context of others—as for instance

the sort of holism inferentialism involves has many more resources available for explaining what it is for two interlocutors to be applying the *same* concept when concepts are understood in the Kantian way, in terms of their normative grip on us, rather than in the Cartesian way, in terms of our epistemic grip on them. One thread that runs through much of the tapestry I've been displaying is the way in which an inferentialist semantic metalanguage gives us the resources to explain what is expressed by the locutions that make explicit specifically *representational* relations. How one understands the relation between the expressive and the representational dimensions of intentionality—between 'that' intentionality and 'of' intentionality—must be at the core of any theory of content or intentionality.

NOTES

1. In *A Study of Concepts* (Cambridge, MA: Bradford Books/MIT Press, 1995).
2. At the level of generality in play in this telling of the story. Of course, there is *lots* more going on in Lewis than this—but Fodor is not *just* an empiricist either.
3. In H.-N. Castañeda, ed. *Action, Knowledge, and Reality* (Indianapolis: Bobbs-Merrill, 1975), p. 285.
4. In fact, it is also an essential move to replace the undifferentiated notion of propriety with the more articulated distinction of two flavors of deontic status: commitment and entitlement. Doing that permits one to define incompatibility (two claims are incompatible if commitment to one precludes entitlement to the other), and three sorts of inferential relations: commitment-preserving inferences (a generalization to the case of material inferences of *deductively* good inferences), entitlement-preserving inferences (a generalization to the case of material inferences of *inductively* good inferences), and incompatibility entailments (a generalization to the case of material inferences of *modally robust*, i.e., counterfactual-supporting, inferences). See Chapter 6 of *Articulating Reasons* (Cambridge, MA: Harvard University Press, 2000) for a rationale.
5. J. Fodor and E. Lepore, "Brandom's Burdens: Compositionality and Inferentialism," *Philosophy and Phenomenological Research* LXIII, no. 2 (September 2001): pp. 465–83.
6. Joe Camp's instructive example in *Confusion* (Cambridge, MA: Harvard University Press, 2002), p. 14.
7. I take it that this is how Fodor and Lepore see things, though their discussion is not as clear on this point as one would like. They say 'Like everybody else who thinks that content comes from inferential role, Brandom needs, but doesn't have, a story about which inferences are the ones that content comes from.' ("Brandom's Burdens," p. 471) As I point out in the text, Sellars does have such a story. But everyone needs to make the distinction only if it is incoherent to appeal to all the inferences, and so to refuse to make the supposedly required distinction between 'what the world contributes to the reliability of inferences and what the (putative) rules of language contribute.' Acknowledging in a footnote that I am 'explicitly suspicious' of the distinction, the authors say: 'So be it; but then he needs some argument that he can make sense of the notion of content without employing it. To say that he has no such argument on offer would be to put the case very mildly.' Working out a theory of that shape, as I do in Part Two of *Making It Explicit* (Cambridge, MA: Harvard University Press, 1994), ought to be the very best such argument. But in fact Fodor and Lepore seem to have in mind their worries about holism, which I address in the next section. Insofar as the holism is the source of the problem with not privileging some inferences, there is no *separate* issue about 'which inferences.' I address the holism question in the next section, where I sketch some of the arguments that are given in more detail in *Making It Explicit*.
8. As Israel Scheffler and Paul Feyerabend had pointed out already in the middle-1960s.
9. I do think there are some deep issues about holism in the vicinity. I discuss some of them in Chapter 6 of *Tales of the Mighty Dead: Historical Essays in the Metaphysics of Intentionality* (Cambridge, MA: Harvard University Press, 2002).
10. "Brandom's Burdens," op. cit., p. 480.
11. See my "Singular Terms and Sentential Sign Designs," *Philosophical Topics* XV, no. 1 (Spring 1987): pp. 125–67; and Bas Van Fraassen's "Quantification as an Act of Mind," *Journal of Philosophical Logic* 11, no. 2 (August 1982): pp. 343–69.
12. Of course, not everyone agrees that this construction is successful. See Alex Oliver's *Times Literary Supplement* (*TLS*) review of *Articulating Reasons*, and J. Fodor and E. Lepore. They are quite right to point out that the characterizations of singular term and predicate will work for natural languages only if they are first somewhat regimented or syntactically pre-processed. It is not hard to show how to do that so as to deal with the cases they forward as counterexamples. (So, for instance, Oliver's case of the non-substitutability of 'the first Postmaster General' for 'Ben Franklin' in the phrase 'good old Ben Franklin' requires that appositives be made explicit, in a form such as 'Ben Franklin, who is good and old,' in which case the substitution is syntactically allowed and semantically correct. Cf. *Making It Explicit*, p. 388, where this construction is discussed.) Transformations need to be applied before categorial analysis, and after categorial synthesis. This is a point, again, that is explicitly acknowledged in *Making It Explicit*, even though it is not addressing that phase of the process.

    Another sort of criticism both sets of critics make, however, is based on a simple misunderstanding of the view they are criticizing. Thus, to pick a representative example from Fodor and Lepore (p. 476), they object to the denial that there can be systematically asymmetric substitution relations among singular terms, as there can be among predicates, that 'Father

was at Magdalen,' entails 'Father was at Oxford,' but not vice versa. But this looks like a counterexample only if one drops the crucial initial quantifier from the claim. For there to be an asymmetric relation of the kind asserted, it must be the case that the inference from P(Magdalen) to P(Oxford), but not the converse, holds for *every* predicate P, not just for some specific one. And this is obviously not the case: consider P=Peter has never set foot in . . . (Of course, one's next thought would be that, so understood, the corresponding asymmetry does not hold for the predicates . . . walks and . . . moves. But that is not right. That thought overlooks the fact that the predicates in question are not *simple* predicates, but *complex* ones—in Dummett's vocabulary. That is, they are sentence frames identifiable with equivalence classes of sentences that are substitutional variants of one another.) All this is explicitly discussed, along with other putative counterexamples, in the extended section VI of Chapter Six of *Making It Explicit*, "Objections and Replies," to which the reader of *Articulating Reasons* was referred for the details of the argument.

13. And is, in the Appendix I to Chapter 6 of *Making It Explicit*. The Lewis article is in G. Harman and D. Davidson, *Semantics for Natural Languages* (Dordrecht: Reidel, 1972).
14. J. Fodor and E. Lepore, op. cit., p. 472.

The Intentionality of Phenomenology and the Phenomenology of Intentionality

Terence Horgan and John Tienson[1]

What is the relationship between phenomenology and intentionality? A common picture in recent philosophy of mind has been that the phenomenal aspects and the intentional aspects of mentality are independent of one another. According to this view, the phenomenal character of certain mental states or processes—states for which there is 'something it is like' to undergo them—is not intentional. Examples that are typically given of states with inherent phenomenal character are sensations, such as pains, itches, and color sensations. This view also asserts, on the other hand, that the intentionality of certain mental states and processes—their being about something—is not phenomenal. Beliefs and desires are the paradigm cases of intentional mental states. Although they are intentionally directed—i.e., they have aboutness—these mental states are not inherently phenomenal. There is nothing that it is like to be in such a state by virtue of which it is directed toward what it is about.

We will call this picture *separatism*, because it treats phenomenal aspects of mentality and intentional aspects of mentality as mutually independent, and thus separable. Although there may be complex states that are both phenomenal and intentional, their phenomenal aspects and their intentional aspects are separable. Many philosophers who hold this picture have thought that these two aspects of mentality lead to quite different sorts of problems with respect to the project of 'naturalizing the mental.' Proponents of separatism often hold that while the problem of naturalizing phenomenology poses great difficulties, the problem of naturalizing intentionality is much more tractable.[2]

Separatism has been very popular in philosophy of mind in recent decades, and is still widely held. Those who oppose it regard it as a view against which they need to characterize their own positions—a common picture that they must explicitly reject. In this paper we argue that separatism is profoundly wrong. We depart from it quite thoroughly, in ways importantly different from other recent departures. We affirm the following theses, both of which are repudiated by separatism:

> *The Intentionality of Phenomenology:* Mental states of the sort commonly cited as paradigmatically phenomenal (e.g., sensory-experiential states such as color-experiences, itches, and smells) have intentional content that is inseparable from their phenomenal character.

First published in the first edition of this book: D. J. Chalmers, ed., *Philosophy of Mind: Classical and Contemporary Readings* (Oxford University Press, 2002), pp. 520–33.

The Phenomenology of Intentionality: Mental states of the sort commonly cited as paradigmatically intentional (e.g., cognitive states such as beliefs, and conative states such as desires), when conscious, have phenomenal character that is inseparable from their intentional content.

In addition to these two theses (henceforth, IP and PI), we advocate another important claim about the interpenetration of phenomenology and intentionality:

Phenomenal Intentionality: There is a kind of intentionality, pervasive in human mental life, that is constitutively determined by phenomenology alone.

We use the expression 'constitutively determined' to mean that this kind of intentionality is not merely nomically determined; rather, intentional mental states have such intentional content *by virtue of* their phenomenology.

So-called 'representationalist' theories of phenomenal properties are a currently influential departure from separatism.[3] Although extant versions of representationalism embrace thesis IP, typically they do not embrace thesis PI. Nor do they embrace the thesis of Phenomenal Intentionality, since they hold that intentionality is prior to phenomenology. So our position differs significantly from standard representationalism.[4]

We argue for the three theses set out above (sections 1–3), in part by way of introspective description of actual human experience. If you pay attention to your own experience, we think you will come to appreciate their truth.[5] Our position has important consequences, when combined with the plausible thesis (argued for in section 4) that phenomenology is 'narrow,' i.e., it does not depend—except perhaps causally—upon what goes on 'outside the head' of the experiencer. One consequence is that there is a kind of *narrow* intentionality that is pervasive in human mental life—a form of intentional directedness that is built into phenomenology itself, and that is not constitutively dependent on any extrinsic relations between phenomenal character and the experiencer's actual external environment. A further consequence is that theories that ground all intentionality in connections to the external world, such as causal and teleological theories of intentionality, are deeply mistaken.

1. The Intentionality of Phenomenology

The mental states typically cited as paradigmatically phenomenal have intentional content that is inseparable from their phenomenal character. Let us consider some examples: first, experiences of red as we actually have them. You might see, say, a red pen on a nearby table, and a chair with red arms and back a bit behind the table. There is certainly something that the red that you see is like to you.[6] But the red that you see is seen, first, as a property of objects. These objects are seen as located in space relative to your center of visual awareness. And they are experienced as part of a complete three-dimensional scene—not just a pen with table and chair, but a pen, table, and chair in a room with floor, walls, ceiling, and windows. This spatial character is built into the phenomenology of the experience.

Consider too the experience of seeing an apple on the table, picking up the apple, and taking a bite out of it. There is the look and smell of the apple. Then (as you grasp it) there is the feel of the apple, its smoothness, roundishness, and firmness. Then there is its weight (as you pick it up). Finally there is the feel of the apple in your mouth, followed by the crunching sound, taste, and feel of juiciness as you take a bite. We will not attempt to write the small book one could write describing this simple experience. But we need to note some highlights. First, the look, feel, smell, sound, and taste of the apple are experienced as a unity in space, as all belonging to a single object. The taste is in your mouth; the smoothness and roundishness that you feel—with parts of your mouth as well as your hand—are there, too. Second, it is important to notice that what is experienced tactilely are various spatial properties *of the object*, not sensations. One has, of course, tactile sensations as well, though one does not normally attend to them. (The tactile sensations are, when noticed, experienced as the sort of things that can only belong to a sentient being.) The properties of smoothness, firmness, etc. are experienced as the sorts of things that can only belong to an 'external' object in space.[7] Third, the apple is encountered as *moving*. The experience is of a *temporal* object, an object that endures. The same is true when you see another person take a bite of an apple. Experience is not of instants; experience is temporally thick. This is obvious in the case of hearing tunes or sentences, where the temporal *pattern* is a palpable feature of

the experience. The temporal pattern is also a palpable feature of the seen moving apple, though less frequently noted as such.[8] But it is no less true that stationary objects are *seen as* enduring and as unchanging.[9]

For any experience involving a specific shade of red, one can abstract away from the total experience and focus on the distinctive what-it's-like of that shade of red per se—a phenomenal aspect of this total experience that it has in common with innumerable other total experiences that differ in the perceived location of the experienced red or in the shape of the red surface, etc. But even considered in isolation from any total visual-experiential state, the what-it's-like of experiencing red is already intentional, because it involves red *as the intentional object of one's experience.* Again, redness is not experienced as an introspectible property of one's own experiential state, but rather as a property of visually presented objects.[10]

Of course, in typical cases of experiencing red, the *overall* phenomenal character of one's visual experience is a structurally rich what-it's-like of *experiencing a visually presented scene*, a scene that contains a whole array of apparent enduring objects with various properties and relations—including the property redness instantiated on the surfaces of some of these objects. The total visual experience with this overall phenomenal character is richly intentional, since it presents a temporally extended scene comprising various objects that instantiate various properties and relations at various spatial locations relative to one's center of visual awareness. This total visual experience is also richly *phenomenal*, because there is an overall what-it's-like of experiencing the whole scene. (Any visually noticeable alteration in the visually presented scene would be a *phenomenal* difference in one's total visual experience.)

Another commonly cited example of phenomenal consciousness is the distinctive phenomenal character of pain. Experiences of pain are thoroughly intentional: pain is experienced as a particular feeling *at a certain place in one's own body.* (This is why there can be such a thing as phantom-limb phenomena, in which pain is experienced as located in a limb that has been amputated.)

More generally, the overall phenomenal character of one's experience includes a structurally rich what-it's-like of *tactilely and kinesthetically experiencing one's presented body*, an apparent body containing a whole array of tactilely and kinesthetically distinguishable apparent parts, many of which are experienced as parts that one can voluntarily move. The total tactile/kinesthetic experience with this overall phenomenal character is richly intentional, by virtue of the complexity of the body as presented. This total experience is also richly *phenomenal*, since it has the what-it's-like-ness of tactilely and kinesthetically experiencing one's whole body. (Any tactilely or kinesthetically distinguishable alteration would be a *phenomenal* difference in one's total tactile/kinesthetic experience.)

The full-fledged phenomenal character of sensory experience is an extraordinarily rich synthetic unity that involves complex, richly intentional, total phenomenal characters of visual-mode phenomenology, tactile-mode phenomenology, kinesthetic body-control phenomenology, auditory and olfactory phenomenology, and so forth—each of which can be abstracted (more or less) from the total experience to be the focus of attention. This overall phenomenal character is thoroughly and essentially intentional. It is the what-it's-like of being an embodied agent in an ambient environment—in short, the what-it's-like of being in a world.

2. The Phenomenology of Intentionality

We have been describing the intentionality of sensory-perceptual phenomenal consciousness.[11] Let us now focus on the thesis of the phenomenology of intentionality (PI): consciously occurring intentional states have phenomenal character that is inseparable from their intentional content.

Intentional states have a phenomenal character, and this phenomenal character is precisely the what-it-is-like of experiencing a specific propositional-attitude type vis-à-vis a specific intentional content. Change either the attitude-type (believing, desiring, wondering, hoping, etc.) or the particular intentional content, and the phenomenal character thereby changes too.[12] Eliminate the intentional state, and the phenomenal character is thereby eliminated too. This particular phenomenal character could not be present in experience in the absence of that intentional state itself.

Consider, for example, an occurrent thought about something that is not perceptually presented, e.g., a thought that rabbits have tails. Quine notwithstanding, it seems plainly false—and

false for phenomenological reasons—that there is indeterminacy as to whether one is having a thought that rabbits have tails or whether one is instead having a thought that (say) collections of undetached rabbit parts have tail-subsets. It is false because there is something that it is like to have the occurrent thought that rabbits have tails, and what it is like is different from what it would be like to have the occurrent thought that collections of undetached rabbit parts have tail-subsets.[13]

The overall phenomenology of these kinds of intentional states involves abstractable aspects which themselves are distinctively phenomenological. For example, if one contrasts wondering whether rabbits have tails with thinking that rabbits have tails, one realizes that there is something *common* phenomenologically—something that remains the same in consciousness when one passes from, say, believing that rabbits have tails to wondering whether rabbits have tails, or vice versa. It is the distinctive phenomenal character of holding before one's mind the content *rabbits have tails*, apart from the particular attitude type—be it, say, wondering, hoping, or believing. This aspect of the overall phenomenology of intentionality is *the phenomenology of intentional content.*[14]

In addition, there is also a specific what-its-likeness that goes with the attitude type as such. There is a phenomenological *difference* between wondering whether rabbits have tails on one hand and thinking that rabbits have tails on the other. This aspect is *the phenomenology of attitude type.* Attentive introspection reveals that both the phenomenology of intentional content and the phenomenology of attitude type are phenomenal aspects of experience, aspects that you cannot miss if you simply pay attention.

One might reply that although there is indeed a phenomenological difference between thinking that rabbits have tails and thinking that collections of undetached rabbit parts have tail-subsets, this difference merely involves the fact that we often think in language. Thus, the phenomenological difference between the two thinking experiences involves not the different contents, but rather the fact that the auditory imagery that goes with thinking that rabbits have tails is different from the imagery that goes with thinking that collections of undetached rabbit parts have tail-subsets.

However, nonperceptual intentionality in normal humans does not always involve language and/or auditory imagery. For instance, conscious, unverbalized beliefs about the locations of nearby unperceived objects are just as ubiquitous in human life as is the explicit or imagistic verbalization of one's focal train of thought. Think for example, of cooking, cleaning house, or working in a garage or wood shop. In any such activity, you might spontaneously move to retrieve a needed tool that is out of sight. There is something that it is like to think that a certain tool is just there—in that cabinet, say—but such beliefs are typically not verbalized either vocally or subvocally or by way of verbal imagery. (Your verbal energies might all the while be directed toward ongoing philosophical discussion with a companion, uninterrupted by your selection of an appropriate tool.) You also, of course, frequently have unverbalized thoughts about the locations of objects in distant familiar locations.

In any event, the what-it's-likeness of intentionality that we are talking about—even when it is specifically tied to certain words in English or some other natural language—does not attach to those words simply as sequences or patterns of sounds, or even as syntactic structures. It attaches to awareness of those words qua contentful; it is the what-it's-like of hearing or saying those words when they mean just that: that rabbits have tails. So the basic point holds: even if thinking did always involve auditory imagery, the auditory imagery would be intentionally loaded in the experience, not intentionally empty.

This point is illustrated and defended by Galen Strawson 1994. Strawson discusses what he calls 'understanding experience.' He contends that understanding and other related kinds of occurrent intentional mental states and processes are very commonly, if not always, laden with distinctive what-it's-likeness. He points out, for example, the phenomenological difference between hearing speech in a language that one does not understand and hearing speech in a language that one does understand. Imagine two people side by side hearing the same spoken sequence of sounds, with one of them understanding the language and the other one not. At a certain relatively raw sensory level, their auditory experience is phenomenologically the same; the sounds are the same, and in some cases may be experienced in much the same way *qua sounds.* Yet it is obvious introspectively that there is something phenomenologically very different about what it is like for each of them: one person is having understanding experience with the distinctive phenomenology of understanding the sentence to mean just what it does, and the other is not.

Consider, as a similar example for a single speaker, first hearing 'Dogs dogs dog dog dogs,' without realizing that it is an English sentence, and then hearing it as the sentence of English that it is. The phenomenal difference between the experiences is palpable. (If you do not grasp the sentencehood of the 'dogs' sentence, recall that 'dog' is a verb in English, and compare, 'Cats dogs chase catch mice.')

Consider also hearing an ambiguous sentence. One typically hears it as meaning some one thing in particular, often without realizing that it is ambiguous. There is a phenomenological difference, for example, between hearing 'Visiting relatives can be boring,' as a remark about the people who are visiting, vs. hearing it as a remark about visiting certain people oneself. Again, imagine hearing or saying 'Time flies' as a cliché about the passage of time, vs. saying or hearing it as a command at the insect races. The actual sound or auditory imagery may be the same, but the total experiences are phenomenally quite different. The sound may have some role that would make it appropriate to call it a *vehicle* of intentionality, but its meaning what it does, having the intentional content that it does, is an entirely different aspect of the overall phenomenal character of the experience.

In sum: Cognitive intentional states such as consciously occurrent thoughts, and conative intentional states such as consciously occurrent wishes, are phenomenal qua intentional. The overall phenomenal character of such a state comprises both the phenomenology of its specific intentional content and the phenomenology of its specific attitude-type. These are abstractable phenomenal aspects of the state's unified phenomenal character: the what-it's-like of undergoing this specific propositional attitude vis-à-vis that specific intentional content.

3. Phenomenal Intentionality

The intuitive considerations in the last two sections can be developed into an argument for the thesis of phenomenal intentionality: there is a pervasive kind of intentionality that is constitutively determined by phenomenology alone. One way to articulate and sharpen this claim is the following. Let two creatures be *phenomenal duplicates* just in case each creature's total experience, throughout its existence, is phenomenally exactly similar to the other's. We can then state the Phenomenal Intentionality thesis this way:

> There is a kind of intentional content, pervasive in human mental life, such that any two possible phenomenal duplicates have exactly similar intentional states vis-à-vis such content.

We will call this type of content *phenomenal* intentional content. The full range of a creature's phenomenal intentional content is determined and constituted *wholly by phenomenology.*

Consider any creature who is a complete phenomenal duplicate of yourself—its mental life is phenomenally exactly like yours. Assume nothing *else* about this creature. The thought experiment thus builds in an epistemic 'veil of ignorance' about this creature, in order to filter out any factors other than phenomenology itself. So for all you know about this arbitrary phenomenal duplicate of yourself, its sensory-perceptual experiences—including its tactile-kinesthetic experiences and its embodied-agency experiences—might be very largely illusory and hallucinatory concerning the real nature of itself and its surroundings. (It may be helpful to imagine the phenomenal duplicate as a brain in a vat or a disembodied Cartesian mind.) We will argue that you and your phenomenal duplicate share a pervasive kind of intentional content—phenomenal intentional content.[15]

As argued in section 1, sensory-phenomenal states and processes have intentional content that is inseparable from their phenomenal character. These states present an apparent world full of apparent objects that apparently instantiate a wide range of properties and relations, and they present oneself as an apparently embodied agent within that apparent world. Since this kind of intentionality is inseparable from phenomenal character, your phenomenal duplicate will have an apparent world presented to it in exactly the same way.

To make the general point with a representative example, suppose that you have the experience of seeing a picture hanging crooked. Each of your phenomenal duplicates has a phenomenally identical experience. Some of these experiences will be accurate and some will be inaccurate. Whether or not a given duplicate's picture-hanging-crooked experience is accurate—that is, whether or not things are as the experience presents things as being—will depend upon the duplicate's actual environment. Thus, the sensory-phenomenal experience, by itself, determines conditions of accuracy: i.e., a class of ways the environment must be in order for the experience to be accurate.[16] In order for

such an experience to be accurate, there must be a picture before oneself, and it must be crooked.

That these phenomenally identical experiences all have the same truth conditions is reflected in the fact that each of the experiences is subject in the same way to investigation as to whether it is accurate.[17] For example, you and your phenomenal duplicate each might have the experience of seeming to oneself to be testing one's perceptual experience for accuracy by making measurements or using a level. You and your phenomenal duplicate each might have the subsequent experience of seeming to oneself to discover that the picture merely appears to be crooked because of irregularities of the wall, or tricks of light. Or, you and your phenomenal duplicate might, in the course of seeming to oneself to be attempting to perform these tests, have the experience of seeming to discover that there actually is no picture—say, by seeming to oneself to discover that one has been looking at a clever holographic image cooked up to make it appear that there is a picture hanging on the wall.[18]

There is also, of course, a sense in which these crooked-picture perceptions of you and your duplicate have different truth conditions. You and your duplicate are looking at different pictures. So the accuracy of your own perception depends on the specific picture you yourself are seeing, whereas the accuracy of your duplicate's perception depends on the specific, and different, picture that it is seeing. There are thus two ways of thinking of truth conditions: as determined wholly by phenomenology, and as determined in part by items in the experiencer's environment that satisfy the experiencer's phenomenology. We return to this distinction in section 5.2.

Your phenomenal duplicate accepts the presentations delivered by perceptual experience—accepts, for example, that there is a picture and a wall—just as you do. These 'belief-wise' acceptance states have exactly the same phenomenology, the what-it-is-like of occurrently believing that thus-and-so (where one's occurrent sensory experience presents things as being thus-and-so). The states also are phenomenologically integrated with those ongoing, richly intentional, sensory-perceptual experiences in exactly the same way as yours. Thus, they are experienced as having the same belief-specific role: the same apparent input conditions, involving apparent deliverances of the apparent body's apparent senses, and the same apparent effects, involving experiences of *apparently acting appropriately* with regard to the apparent world as presented. It seems intuitively clear that a belief-wise acceptance state with these phenomenological features is a genuine belief. The phenomenal character of these states, which includes the phenomenology of role, constitutively determines that they are genuine beliefs.[19] And as we argued above, the sensory-presentational content of these states is the same for you and your phenomenal duplicate.

So far we have been discussing perceptual experience and perceptual belief. But since the phenomenal consciousness of your phenomenal duplicate would provide very rich perceptual presentations of a world of numerous apparent objects instantiating numerous apparent properties and relations, such presented items would thereby figure in a wide range of propositional-attitude states whose content goes well beyond the presentations of perceptual experience itself. Here the phenomenology of intentionality—the what-it's-like of occurrent propositional attitudes as such—enters in full force, quite apart from the presentational content of one's current sensory-perceptual experience. For your phenomenal duplicate, no less than for you yourself, there would be something that it's like to wonder whether to cook meatloaf for dinner, something that it's like to have the thought that there's beer in the fridge, something that it's like to hope that one's spouse isn't angry that one is coming home late from the Philosophy Department party. These occurrent states in the phenomenal duplicate have all the same 'propositional attitude-ish' phenomenology as they do in you. They are experienced as having exactly the same causal role vis-à-vis the phenomenal duplicate's apparent embodied behavior in its apparent world as you experience them as having. It seems intuitively clear that states with all these features qualify as full-fledged *propositional attitudes* in your phenomenal duplicate, just as they do in you.

In addition, for each such propositional-attitude state in yourself and in your phenomenal duplicate, the two states have the same phenomenal intentional content, i.e., the same phenomenologically determined truth conditions. Consider, for example, two phenomenologically identical belief-states that you and your phenomenal duplicate would both express by the string of words 'The picture behind me is crooked.' In order for such a belief to be true, there must indeed be an object in a certain relation to oneself—behind, no intervening walls, etc.—that satisfies the

phenomenologically determined criteria for being a picture, and that picture must be hanging crooked. Considered in this way, your belief and that of your phenomenological duplicate have exactly the same truth conditions.[20] These occurrent states in the phenomenal duplicate, by virtue of having the same phenomenologically determined truth conditions as yours, are thereby subject to the same methods of accuracy assessment: for instance, you and your phenomenological duplicate might each experience turning around to see if the picture is still crooked. If it still appears crooked, you might then experience going through the tests mentioned above. The possibility of such tests is in some sense understood, if not explicitly phenomenologically given, in having the conscious belief that there is a picture hanging crooked behind oneself.

Since your phenomenal duplicate shares with you all the phenomenal intentionality so far described, it thereby possesses significant conceptual resources to speculate and theorize—for instance, about what is very distant spatiotemporally, about what is very small, about the underlying causes of experience and of the apparent ambient environment. It can reason causally, form abstract theoretical concepts, formulate scientific hypotheses and theories, and experience itself as an apparent embodied agent actively engaged (in apparent cooperation with other apparent embodied agents) in the apparent empirical corroboration or disconfirmation of such hypotheses and theories. It can have experiences as of apparently reading about such matters or apparently hearing lectures about them, and thereby can acquire a body of well warranted scientific beliefs about itself and its world. In these respects too, your phenomenal duplicate is like yourself, even though the experiential basis upon which it bootstraps its way up to well warranted, semantically evaluable, scientific beliefs might be highly (or even completely) non-veridical. Thus, for each of your propositional-attitude states about such theoretical entities, your phenomenal duplicate has a propositional-attitude state of the same kind. And for each pair of corresponding states in you and your duplicate respectively, the two states have the same phenomenal intentional content—i.e., the same phenomenally determined truth conditions, linked via the same phenomenally determined 'web of belief' to the same kinds of experiential methods of accuracy-assessment.[21]

Virtually everything we have been saying is really just attentive phenomenological description, just saying what the what-it's-like of experience *is* like. It is just a matter of introspectively attending to the phenomenal character of one's own experience. You and your phenomenal duplicate share a pervasive kind of mental intentionality—viz., phenomenal intentionality.

We take it that this thought-experimental argument supports the idea that each *specific* occurrent intentional state with phenomenal intentional content is constitutively determined by its own distinctive phenomenal character—viz., the what-it's-like of undergoing that particular attitude-type vis-à-vis that particular phenomenal intentional content. That is, specific phenomenal character determines specific intentional states, provided that the experiencing creature has a sufficiently rich network of actual and possible phenomenal/intentional states. Suppose, for example, that you are now undergoing a psychological state with the distinctive phenomenal what-it's-like of believing that a picture is hanging crooked on a wall directly behind you. Then you thereby *believe* that there is a picture hanging crooked on a wall directly behind you; undergoing this phenomenology constitutively determines that you are instantiating that belief-state. Any experiencing creature undergoing this phenomenology would thereby instantiate the belief-state, even if its overall phenomenology is otherwise quite different from your own.

Although each occurrent intentional state with phenomenal intentional content is constitutively determined by its own phenomenal character, at least in the context of a full-fledged cognitive system, it is important to appreciate that this does not mean that phenomenal intentionality somehow guarantees infallible knowledge about what one's first-order intentional states are. Beliefs about one's own intentional states are *second-order* intentional states, and the Phenomenal Intentionality thesis is compatible with the possibility that such beliefs are sometimes mistaken. Indeed, the thesis is compatible with the possibility that some creatures who have phenomenal intentionality—say, certain kinds of nonlinguistic animals—entirely lack any capacity to undergo second-order intentional states at all. What-it's-likeness is one thing; discursive judgments *about* it are another. Such judgments are fallible (as are judgments about most anything), even though humans do possess especially reliable capacities to form accurate introspective discursive/classificatory judgments about their own phenomenology.[22]

4. The Narrowness of Phenomenology and of Phenomenal Intentionality

Phenomenology does not depend constitutively on factors outside the brain. Now, it is obvious enough that in normal humans, phenomenology does depend *causally* on some such factors; but one need only consider how this causal dependence works in order to appreciate the lack of constitutive dependence. First, phenomenology depends causally on factors in the ambient environment that figure as *distal* causes of one's ongoing sensory experience. But second, these distal environmental causes generate experiential effects only by generating more immediate links in the causal chains between themselves and experience, viz., physical stimulations in the body's sensory receptors—in eyes, ears, tongue, surface of the body, and so forth. And third, these states and processes causally generate experiential effects only by generating still more immediate links in the causal chains between themselves and experience—viz., afferent neural impulses, resulting from transduction at the sites of the sensory receptors on the body. Your mental intercourse with the world is mediated by sensory and motor transducers at the periphery of your central nervous system. Your conscious experience would be phenomenally just the same even if the transducer-external causes and effects of your brain's afferent and efferent neural activity were radically different from what they actually are—for instance, even if you were a Brain in a Vat with no body at all, and hence no bodily sense organs whose physical stimulations get transduced into afferent neural inputs.[23] Among your logically possible phenomenal duplicates, then, are beings whose sensory experience is radically illusory, in the manner of the famous Evil Deceiver scenario in Descartes' First Meditation—or its contemporary version, the Brain in a Vat.

Thus, phenomenology is *narrow*, in the sense that it does not depend constitutively on what's outside the skin, or indeed on what's outside of the brain. We can now make the central argument:

(1) There is pervasive intentional content that constitutively depends on phenomenology alone.

(2) Phenomenology constitutively depends only on narrow factors.

So, (3) There is pervasive intentional content that constitutively depends only on narrow factors.

That is, the theses of phenomenal intentionality and the narrowness of phenomenology jointly entail that there is kind of *narrow* intentional content (the kind we have dubbed *phenomenal* intentional content), pervasive in human life, such that any two creatures who are phenomenal duplicates must also have exactly similar intentional states vis-à-vis this kind of narrow content. Phenomenal intentional content is pervasive and narrow. Any adequate philosophical and scientific conception of mind should accommodate this conclusion.[24]

5. Some Philosophical Morals

We now draw some morals from the preceding discussion, first about strong externalist theories of intentionality, second about how phenomenal intentional content is related to mental reference and to wide content, and third about the extent of the so-called 'hard problem' of phenomenal consciousness.

5.1. Strong Externalist Theories of Mental Intentionality Are Wrong

We certainly do not deny that there is such a thing as 'wide content' in language and in thought. Important lessons have been learned from Kripke, Putnam, Burge, and others about the relevance of the external environment in contributing both to the meaning of certain terms in natural language (e.g., natural-kind terms like 'water') and to certain aspects of the content of thought (e.g., aspects of thought that one would express verbally by employing the term 'water'). But Putnam's famous slogan that 'meaning ain't in the head' is properly construed as asserting that *not all* meaning is in the head; it doesn't begin to follow from this, or from the considerations adduced in support of it, that *no* meaning is in the head. We will return to the topic of wide content presently.

However, a number of current theories of mental intentionality are *strongly* externalist: they assert that all intentionality is grounded in causal connections between states of the cognitive system and states of the external world; there can be no intentionality without some suitable kind of actual connection between what is going on in the head and the wider environment. Strong externalist theories of intentionality include (i) causal theories of content that find the necessary connection in the causal antecedents of the state, (ii) covariational

theories that find the connection in certain kinds of systematic correlations between occurrences of an internal state and occurrences of an external state of affairs, (iii) teleosemantic theories that look to environmentally situated proper functions that certain internal states possess in virtue of evolutionary design, and (iv) learning-based theories that invoke internal adaptational changes in the creature's own history.[25]

Given our conclusions in sections 1–4, it follows that strong externalist theories of intentionality are wrong. They are not just slightly wrong, not just wrong in detail. Rather, they are fundamentally mistaken, because they claim to naturalize the entire phenomenon of mental intentionality and yet there is a rich and pervasive kind of narrow intentionality—phenomenal intentionality—that is constitutively independent of external factors. Strong externalist theories therefore cannot successfully naturalize the full phenomenon of mental intentionality, because they utterly fail to aim at one crucial aspect of it. Ideas employed by strong externalists might still have a useful role to play, however, in contributing to philosophical understanding of phenomena like wide content and mental reference, topics to which we now turn.

5.2. Phenomenal Intentionality, Mental Reference, and Wide Content

Suppose Alfred and Bertrand are looking at two different barns, and each of them says, 'That's an old barn.' Do their statements have the same truth conditions? Yes and no. In one way, they have different truth conditions. Alfred's statement is made true or false by the age of the barn that he is looking at, while Bertrand's statement is made true or false by the age of the distinct barn that *he* is looking at. Following recent usage, we will call such truth conditions, which depend on the actual entities referred to in a statement or thought, *wide truth conditions*. But in another way, Alfred's and Bertrand's statements have the same truth conditions. In each case the truth condition is that there must be an actual barn that he is looking at (and not, for example a papier-mache mock-up of a barn, or only the facing side of a stage 'barn' on a movie set), and that barn must be old. Such truth conditions are *narrow truth conditions*. They are determined skin-in, so to speak, and are completely determined by phenomenology. In our view, the situation is similar with respect to phenomenologically identical intentional states shared by phenomenal duplicates.

In section 3 we discussed belief-wise acceptance of the deliverances of perceptual experience. Such acceptance is the normal, default attitude. But it can be cancelled. If you have a lump on a finger, then objects that are smooth and flat will feel as though they have a lump where that finger touches them. But you soon learn not to believe that the object is lumpy. There is similar phenomenology of acceptance concerning propositional attitudes. There is a relevant phenomenal difference, for instance, between these two states: (i) believing that Bill Clinton was U.S. President, and (ii) the state you are in when you say (without believing) that Santa Claus brings presents. The salient difference turns on the fact that the phenomenal character of the first state includes the what-it's-like of accepting the existence of Bill Clinton, whereas the phenomenal character of the second state includes the what-it's-like of believing that Santa Claus does *not* exist. Similarly, suppose you hope or fear that an object of a certain description will be found. There is a clear phenomenal difference between the case in which you know full well that there is such an object and the case in which you do not know whether or not there is such an object.

Phenomenal intentional content presents to consciousness an apparent world that goes far beyond what one is conscious of at any one time; presuming so is itself an aspect of the overall phenomenal character of experience. Phenomenal intentionality thereby determines a complex set of presuppositions concerning the existence of, the persistence of, and various features of, the sorts of entities presented in experience: presuppositions about individuals (including flora, fauna, and other creatures like yourself), kinds, properties, relations, processes, and events of that world. For reasons that will become clear, we call these presuppositions *grounding presuppositions*. They have the phenomenology of acceptance discussed in the previous paragraph. In making a grounding presupposition, one takes it that there *really exists* an entity of a certain sort; and normally, one also presupposes that the (putative) entity in question has certain specific attributes. If there is an actual entity satisfying that presupposition (or satisfying it near enough), then one's thoughts that are intentionally directed toward such a putative entity will *refer* to the actual entity in question; and so the properties of the satisfier will determine whether the beliefs about it are true or false, whether hopes and desires about it are satisfied, and so forth.

Thus, *wide* truth conditions for those beliefs are determined by phenomenal intentionality plus the actual satisfiers of the relevant presuppositions. However, what it takes *to be* a satisfier of the presuppositions is determined by phenomenal intentionality alone. So, when these presuppositions are included in truth conditions, we get *narrow* truth conditions that are thereby determined solely by phenomenal intentionality.[26]

Consider, for example, thoughts about individuals.[27] You, your Twin Earth doppelganger, and your Cartesian duplicate all have certain phenomenologically identical thoughts that you each take to be about a person named 'Bill Clinton.'[28] Hence these thoughts *presuppose* the existence of such a person. Your own thoughts are about the actual Bill Clinton. Your Twin Earth doppelganger's thoughts are about a different person on Twin Earth. You and your Twin Earth doppelganger have thoughts about different individuals, of course, because what a person's thoughts are about—or *refer to*—depends not only on phenomenal intentional content, but also on certain relations between the thinker and the thing the thought is about. Your Cartesian duplicate also has thoughts that purport to be about a person named 'Bill Clinton,' but since the Cartesian duplicate has not been in the right sort of relations to any such person, the Cartesian duplicate's thoughts are not about anyone—they lack reference.[29] *Referring* to something, mentally or linguistically, requires appropriate relations to that thing; but having thoughts that are *intentionally directed* toward such a thing—thoughts *purporting* to refer to such a thing—does not.

Straightforwardly, your thoughts about Bill Clinton are made true or false by facts about Bill Clinton, and your Twin Earth doppelganger's phenomenologically identical thoughts are made true or false by facts about the person who satisfies your duplicate's corresponding presupposition. There is no person who satisfies your Cartesian duplicate's corresponding presupposition, so there is nothing that can be a truth maker for its thought that would be expressed by, say, 'Bill Clinton is a womanizer.'[30]

The differing truth conditions just mentioned are wide truth conditions. But again, there are two ways of thinking about the truth conditions of the phenomenologically identical thoughts of you and your duplicates. In one way, the truth conditions depend upon what is actually referred to (if anything) in those thoughts; this makes them 'wide.' But in another and more fundamental way, the truth conditions are narrow, because what *can be* referred to in those thoughts is determined by phenomenal intentionality—in particular, by the phenomenally given grounding presuppositions. The thought will have wide content only if something in the thinker's environment satisfies the phenomenal intentional grounding presuppositions of that thought. That is, wide content is grounded by phenomenologically determined presuppositions, which are an aspect of phenomenal intentionality and hence are narrow.

As a consequence, the strong externalist theories of intentionality discussed in the previous subsection are not wrong just because they leave out a kind of intentionality—viz, phenomenal intentionality. They are wrong because what they leave out is the *fundamental* kind of intentionality: the narrow, phenomenal kind that is a prerequisite for wide content and wide truth conditions. Because narrow phenomenal content determines wide content, an adequate account of wide content requires a prior account of narrow content.

Because of relevant similarities between singular reference and natural-kind categories, similar observations can be made concerning the narrow and wide truth conditions of thoughts about natural kinds.[31] You, your Twin Earth doppelganger, and your Cartesian duplicate all have phenomenally identical thoughts with the same narrow truth conditions. For all three of you, these thoughts are intentionally directed toward certain small, common furry critters that meow, rub legs, drink milk, etc. For all three of you, these thoughts have the grounding presupposition that there is a natural kind of which these critters are members. But because of differences concerning the satisfiers (if any) of the common grounding presuppositions, these phenomenally identical thoughts have different wide truth conditions. Your own thoughts are made true or false by cats; your Twin Earth doppelganger's phenomenologically identical thoughts are made true or false by cat-like critters of the kind that she or he and others in her or his community have encountered. Suppose that Putnam's story in which cats are spy robots controlled from Mars is true concerning Twin Earth: the critters called 'cats' on Twin Earth are robots controlled from Twin Mars. Then your belief that cats are animals is true; your Twin Earth doppelganger's corresponding belief is false. That is, there are wide truth conditions for these thoughts that are partially determined by features in the environment that may be unknown to the thinker. But again:

these wide truth conditions, differing as they do for your thoughts and your Twin Earth doppelganger's phenomenally identical thoughts, are grounded on shared *narrow* truth conditions.

Your Cartesian duplicate has thoughts that are phenomenally identical to your cat-thoughts, and that have the same narrow truth conditions as yours do. Your Cartesian duplicate's thoughts, like yours, are intentionally directed toward—and thus presuppose—small furry critters of a certain kind. But there are no such critters that the Cartesian duplicate has encountered, directly or indirectly, so there is no kind to which the Cartesian duplicate's thoughts refer. This being so, those thoughts do not have wide truth conditions. So the Cartesian duplicate's thoughts that are phenomenologically identical to your own cat-thoughts lack the kind of wide content that your own thoughts possess and your Twin Earth doppelganger's thoughts also possess.

5.3. The Whole Hard Problem

We are among those who believe that what David Chalmers 1995, 1996 has called 'the hard problem' of phenomenal consciousness is indeed a very hard problem. This is the problem of explaining why it should be that such and such mental state should be *like this*—that is, why it should have the specific what-it's-like aspect it does, rather than having some other phenomenal character or having none at all. Presumably what it is like for you to undergo a particular mental state depends nomologically on what is going on in your brain, inside of the transducers. But why what depends on this brain process should be *like this*, rather than being some other way or being no way at all, seems inexplicable. Standard materialistic treatments of phenomenal consciousness in philosophy and in cognitive science do not close this 'explanatory gap' (as it is dubbed by Joseph Levine 1983); rather, they appear to just leave out the intrinsic what-it's-like aspect of mentality.[32]

In our view, the hard problem is a very pressing and very puzzling conundrum. But its scope is considerably broader than it has often been thought to be. If separatism were correct—i.e., if phenomenology were indeed non-intentional, and intentionality were indeed non-phenomenal—then the hard problem would be limited to the what-it-is-like of non-intentional sensory experience, and would not infect the intentional aspects of mentality. Indeed, discussions of the hard problem often presuppose separatism.[33] But the *whole* hard problem incorporates phenomenal intentionality. Phenomenal consciousness is intentional through and through.

This adds a dimension to the hard problem that often goes unrecognized. Conscious intentional states are intrinsically, *by their very nature*, directed toward whatever they are directed toward.[34] Thus, the hard problem includes this: why should a mental state that is grounded in this physical or physical/functional state be *by its intrinsic phenomenal nature* directed in this precise manner? And this is a very hard problem indeed.[35]

REFERENCES

Addis, L. *Natural Signs: A Theory of Intentionality* (Philadelphia: Temple University Press, 1989).

Brentano, F. *Psychology from an Empirical Standpoint* (London: Routledge, 1973 [original publication 1874, 1924].

Burge, T., "Individualism and the Mental," *Midwest Studies in Philosophy* 4 (1979): pp. 73–121.

Carruthers, P. *Phenomenal Consciousness: A Naturalistic Theory* (Cambridge, UK: Cambridge University Press, 2000).

Chalmers, D. J., "The Puzzle of Conscious Experience," *Scientific American* 273 (1995): pp. 80–86.

_____ *The Conscious Mind* (New York: Oxford University Press, 1996).

_____ "The Components of Content," in *Philosophy of Mind* (New York: Oxford University Press, 2002).

Davies, M., and Humberstone, I., "Two Notions of Necessity," *Philosophical Studies* 38 (1980): pp. 1–30.

Dretske, F. *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995).

Flanagan, O. *Consciousness Reconsidered* (Cambridge, MA: MIT Press, 1992).

Fodor, J. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press, 1987).

_____ "A Theory of Content, II: The Theory," in his *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990). Reprinted in *Mental Representations: A Reader*, Stich, S., and Warfield, T., eds. (Oxford: Blackwell, 1994).

_____ *The Elm and the Expert: Mentalese and Its Semantics* (Cambridge, MA: MIT Press, 1994).

_____ *Concepts: Where Cognitive Science Went Wrong* (Oxford: Oxford University Press, 1998).

Graham, G., and Horgan, T., "Mary Mary, Quite Contrary," *Philosophical Studies* 99 (2000): pp. 59–87.

Horgan, T., and Tienson, J., "Deconstructing New Wave Materialism," in *Physicalism and Its Discontents*, B. Loewer, ed., (Cambridge, UK: Cambridge University Press, 2001).

Husserl, E., "Philosophy as Rigorous Science," in *Phenomenology and the Crisis of Philosophy* (New York: Harper and Row, 1965) (original publication 1910).

Jackson, F. *From Metaphysics to Ethics: A Defense of Conceptual Analysis* (New York: Oxford University Press, 1998).

Kim, J., "Physicalism—or Near Enough," *Princeton Monographs in Philosophy series*, (Princeton: Princeton University Press, 2007).

Kripke, S. *Wittgenstein on Rules and Private Language* (Oxford: Basil Blackwell, 1982).

Lamb, Andrew, "Temporal Dynamics: A Phenomenologically Based Alternative to Four-Dimensionalism and 'Point-Endurantist' Views of Time," *Southern Journal of Philosophy* 39 (2001): pp. 235–59.

Levine, J., "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64 (1983).

Loar, B., "Phenomenal Intentionality as the Basis of Mental Content," in Martin Hahn & B. Ramberg (eds.), *Reflections and Replies: Essays on the Philosophy of Tyler Burge*. (Cambridge: MIT Press, 2003). pp. 229—258.

Lycan, W. *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996).

McGinn, C. *Mental Content* (Oxford: Blackwell, 1989).

_____ *The Problem of Consciousness: Essays Towards a Resolution* (Oxford: Blackwell, 1991).

Reid, T. *Inquiry into the Human Mind on the Principles of Common Sense*, D. Brookes, ed., (University Park, PA: Pennsylvania State University Press, 1997) (original publication 1764).

Ross, J., "Immaterial Aspects of Thought," *Journal of Philosophy* 89 (1992): pp. 136–50.

Siewert, C. *The Significance of Consciousness* (Princeton, NJ: Princeton University Press, 1998).

Silverberg, A., "Narrow Content: A Defense," *Southern Journal of Philosophy* 33 (1995): pp. 109–27.

Stich, S., and Warfield, T., eds. *Mental Representations: A Reader* (Oxford: Blackwell, 1994).

Strawson, G. *Mental Reality* (Cambridge, MA: MIT Press, 1994).

Tienson, J., "An Observation on Common Names and Proper Names," *Analysis* 46 (1986): 73–76.

Tye, M. *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995).

_____ "Phenomenal Consciousness: the Explanatory Gap as a Cognitive Illusion," *Mind* (1999): p. 101.

Van Gelder, T., "Wooden Iron? Husserlian Phenomenology Meets Cognitive Science," in *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, J. Petitot, F. Varela, B. Pachoud, and J. Roy, eds., (Redwood City, CA: Stanford University Press, 1999).

Varela, F., "The Specious Present: A Neurophenomenology of Time Consciousness," in *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, J. Petitot, F. Varela, B. Pachoud, and J. Roy, eds.

NOTES

1. This paper is thoroughly co-authored; order of authorship is alphabetical.
2. For example, Jaegwon Kim (2007) argues, in a way that firmly presupposes the separatist framework, that we can get more than half way to naturalizing the mental, but not all the way. The leading idea is that the intentional aspects of mentality can be naturalized via 'functionalization,' but that the phenomenal aspects resist naturalization.
3. For instance, Dretske 1995; Tye 1995, 1999; Lycan 1996; and Carruthers 2000.
4. Another serious difference is worth mentioning. Representationalists typically regard the problem of naturalizing intentionality as tractable, and they seek to bring phenomenal character within the purview of this putatively tractable naturalization project by construing it as a species of intentionality. We, on the other hand, regard the problem of naturalizing phenomenal character as presently intractable, and we maintain that because of the interpenetration of phenomenology and intentionality, the scope of this intractability includes mental intentionality. Our reasons for holding this view will emerge as this paper progresses.
5. For several recent treatments of the relation between phenomenology and intentionality that are similar in spirit to what we say here, see the discussions of McGinn 1991, Flanagan 1992, Strawson 1994, Siewert 1998, and Loar (2003).
6. This formulation is more accurate that the usual, 'there is something that it is like to see red' because, we take it, the something-that-it-is-like that is referred to is the something-that-it-is-like of the *seen* not of the *seeing*—although, of course, there is something that it is like to see, as opposed, e.g., to hear or imagine.
7. See Thomas Reid, *Inquiry into the Human Mind on the Principles of Common Sense*, Chapter 5.
8. For a recent exception, see Lamb 2001.
9. See Van Gelder 1999 and Varela 1999 for attempts to account for the temporal thickness of experience in cognitive science terms.
10. We think that the right view of these matters is at least very close to that expressed by Laird Addis 1989: 'The idea of the mental is exhausted in all interesting aspects by (1) states of consciousness [primary mental entities]; (2) sensations, emotions, and perception-related entities [secondary mental entities]; and (3) dispositional mental states [tertiary mental entities—i.e., beliefs, etc., that are not presently occurrent]. . . . [O]nly states of consciousness are literally intentional entities. On the other hand, what makes the secondary and tertiary mental entities mental is their relation . . . to states

of consciousness. Secondary mental entities cannot exist except as objects of states of consciousness. . . .' (p. 7). Thus, sensations and sensory qualities exist as, and only as, objects of conscious intentional states. This is essentially Brentano's view in the famous—but widely unread—chapter of *Psychology from an Empirical Standpoint* in which he introduces the word 'intentional' and distinguishes between mental phenomena and physical phenomena. Brentano's mental phenomena are Addis' primary mental entities. Brentano's physical phenomena are Addis' secondary mental entities, not physical things.

11. We do not know exactly when the phrase 'the intentionality of consciousness' first appeared, but one does well to remember that this was *the* phrase that characterized issues concerning intentionality early in the twentieth century.
12. We are talking about psychological changes *discernible to the experiencer*—not changes such as the experiencer's being unknowingly transported to Twin Earth and gradually coming to have 'water'-thoughts about XYZ that are internally indistinguishable from earlier 'water'-thoughts that were about H2O.
13. See Ross 1992 for a nice discussion, congenial to what we are saying here, that focuses not on Quinean indeterminacy but on the 'Kripkenstein' thesis (set forth in Kripke 1982) that there is no fact of the matter whether the symbol '+' means *plus* or *quus*.
14. Part of what makes this aspect of phenomenology essentially a what-it's-like of holding before one's mind a specific *intentional content* is that semantic evaluability is involved: specific *truth* conditions are attached to it. We expand on this important point in section 3.
15. For an excellent discussion that is complementary to what we will say here and that has considerably more detail about phenomenal intentionality, see Siewert 1998, chapters 7 and 8. For another admirable and pertinent discussion, also complementary to ours, see Silverberg 1995; although phenomenological considerations are less prominent in Silverberg's discussion, note his emphasis on how things *seem*, for each member of a group of beings who in effect are phenomenal duplicates respectively inhabiting a variety of different Twin Earthly environments.
16. Siewert 1998 emphasizes the idea that a creature's intentional features are ones for which the creature is assessible for accuracy, and he argues in detail that both perceptual and cognitive experiences are assessable for accuracy by virtue of their phenomenology.
17. We are not here presupposing verificationism, or 'procedural semantics,' or anything of the sort. The point is that differences in sensory-phenomenal content normally are reflected by differences in confirmation/disconfirmation procedures. Thus, sameness of confirmation/disconfirmation procedures provides *strong evidence* for sameness of content (even though it certainly does not *constitute* sameness of content).
18. Of course, even if all these first-person tests for accuracy are successfully passed, this does not guarantee that the sensory-phenomenal experience one is testing really *is* accurate; experiential warrant does not guarantee truth. Also, the reason we talk of *seeming to oneself* to be performing tests and observing outcomes is that a given phenomenal duplicate of yourself might be one whose experiences—including its accuracy-assessment experiences—are systematically non-veridical. This would be so, for instance, for a phenomenal duplicate who is a brain in a vat.
19. There are various further default assumptions involving the intentional objects of perceptual experience that would be psychologically operative in your phenomenal duplicate—normally automatically, as a matter of course—just as they are in you. Examples include default assumptions about experientially presented objects: that these objects have back sides that are not directly presented; that they persist when they are temporarily obscured from view by interposed objects or when one has the experience of looking away from them; that they normally persist when they cease to be presented in experience; and so on.
20. Of course one can also think of the truth conditions of these states as involving the actual, and different, pictures referred to phenomenologically by you and your duplicate, if there are such pictures. There are two kinds of truth conditions for propositional attitudes, just as there are for perceptual states. Truth conditions of one kind are determined phenomenologically, as we have been discussing in the text. And truth conditions of the other kind incorporate the respective objects or kinds (if any) in your own and your phenomenal duplicate's respective ambient environments that are the respective *satisfiers* of the referring concepts in the respective, phenomenologically identical, thoughts. We explore the relationship between these two kinds of truth conditions in section 5.2.
21. The remarks in the preceding footnote apply to these propositional-attitude states too, and to the various components of the relevant web of belief.
22. The fact that there is something epistemically special about first-person introspective access to the phenomenal character of experience is, of course, the basis for the kind of reflective inquiry often called 'phenomenology.' But being epistemically special does not make such introspective judgments *infallible*. For insightful discussion of this complex issue, see chapters 1 and 2 of Siewert 1998.
23. Moreover, phenomenology does not depend constitutively on *anything* beyond phenomenology itself. (Of course, phenomenal character presumably does depend *nomically* on certain states whose nature is describable in non-phenomenological language—in humans, certain brain states.) In this sense, phenomenal character is *intrinsic*. We submit that the intrinsicness of phenomenology (as thus understood) is self-evident to reflective introspection. What-it's-like is what phenomenal consciousness *is*. And what-it's-like is what-it's-like no matter what is going on outside of phenomenology itself.
24. At the scientific level, this means that cognitive science should construe the mind as a 'control system' for effectively operating a potential body in a potential world, regardless of whether or not the control system is actually embodied or en-worlded in the kind of body and world for which its functional architecture is appropriate. This is a common view among cognitive scientists themselves. Such a scientific enterprise is important and tractable, even though it presupposes intentionality rather than explaining it, and even though it does not address the

so-called 'hard problem' of phenomenal consciousness (cf. section 5.3).

25. See Stich and Warfield 1994 for a representative sample of such theories. McGinn 1989 distinguishes two kinds of externalism that he calls 'strong' and 'weak'; he argues against the former, while embracing a teleosemantic approach to the latter.' Although the approach to mental intentionality advocated in Fodor 1987 did acknowledge narrow content, and hence was not a form of strong externalism, it also reflected Fodor's separatism about the phenomenal and intentional aspects of mentality. Because of this, by our lights his construal of narrow content was insufficiently robust, and was a step down the garden path toward strong externalism. He has since gone further down that path; see Fodor 1990, 1994, 1998.
26. Our distinction between narrow and wide truth conditions has some kinship to the approach of so-called two-dimensional modal semantics, which also posits two kinds of truth conditions—one kind narrow and the other kind wide. See Davies and Humberstone 1980; Chalmers 1996, section 2.4, especially pp. 63–65; Jackson 1998, chapters 2 and 3, especially pp. 75–77; and Chalmers (*Philosophy of Mind*, 2002, Chapter 56).
27. Here and below we talk about *thoughts* for ease of exposition, but our remarks will apply to occurrent propositional attitudes in general. If one person doubts what another believes, then their propositional attitudes have the same truth conditions—the truth conditions for what is believed by one and doubted by the other. Similarly for other propositional attitudes.
28. Your Cartesian duplicate is an exact phenomenal duplicate of you that is in the First Meditation situation, thoroughly deluded. You do not really have an *exact* phenomenal duplicate on Twin Earth, however, because on Earth people sometimes have the occurrent thought that water is H2O, whereas on Twin Earth they have instead the thought that water is XYZ. But we will use the useful term 'Twin Earth doppelganger' for a person who is as much like you as is consistent with this difference.
29. It is very common in philosophy of mind to gloss the 'intentional directedness' constitutive of intentionality by saying that intentional states have *aboutness*. (We did so ourselves in the opening paragraph of this paper.) But the word 'about' also is often used to express the relation of *reference*—as we do in the paragraph to which this note is appended, and as we will continue to do below. Both uses can be appropriate in context, but it is important not to conflate them.
30. There is a longstanding dispute over whether in such a case we should say that the thought is false or that it merely lacks a truth value, but this dispute does not affect the issues we are concerned with in this paper.
31. For a discussion of both similarities and differences between singular reference and natural-kind categories, see Tienson 1986.
32. We ourselves have pressed this concern recently; see Graham and Horgan 2000, and Horgan and Tienson (2001).
33. Chalmers himself, however, has not presupposed separatism and has left open the question of which aspects of mentality are by their nature phenomenal—and in particular, whether this is true for intentional states like occurrent beliefs and desires. See, for instance, section 3.3 of Chalmers 1996, especially pp. 19–22.
34. In "Philosophy as Rigorous Science" Husserl criticizes naturalists for holding that intentional states have natures. Conscious intentional states have *essences*, he says; they are essentially directed toward what they are directed toward. This is, in effect, the point we are making. When we say that intentional states are directed *by their very nature* toward what they are directed toward, we do not mean that intentional states have natures in the way in which chemical and physical, and perhaps biological, kinds have natures. Thus, we concur with Husserl's point, although we do not adopt his terminology.
35. We thank William Lycan, Michael Lynch, Brian McLaughlin, Steve Tammelleo, Mark Timmons, and audiences at the University of Arizona and the 2000 Society for Philosophy and Psychology for comments and discussion. Special thanks to David Chalmers and George Graham for their extensive and especially valuable help.

Empiricism and the Philosophy of Mind

Wilfrid Sellars

XI. Thoughts: The Classical View

46. Recent empiricism has been of two minds about the status of *thoughts*. On the one hand, it has resonated to the idea that insofar as there are *episodes* which are thoughts, they are *verbal* or *linguistic* episodes. Clearly, however, even if candid overt verbal behaviors by people who had learned a language *were* thoughts, there are not nearly enough of them to account for all the cases in which it would be argued that a person was thinking. Nor can we plausibly suppose that the remainder is accounted for by those inner episodes which are often very clumsily lumped together under the heading 'verbal imagery.'

On the other hand, they have been tempted to suppose that the *episodes* which are referred to by verbs pertaining to thinking include all forms of 'intelligent behavior,' verbal as well as non-verbal, and that the 'thought episodes' which are supposed to be manifested by these behaviors are not really episodes at all, but rather hypothetical and mongrel hypothetical-categorical facts about these and still other behaviors. This, however, runs into the difficulty that whenever we try to explain what we mean by calling a piece of *nonhabitual* behavior intelligent, we seem to find it necessary to do so in terms of *thinking*. The uncomfortable feeling will not be downed that the dispositional account of thoughts in terms of intelligent behavior is covertly circular.

47. Now the classical tradition claimed that there is a family of episodes, neither overt verbal behavior nor verbal imagery, which are *thoughts*, and that both overt verbal behavior and verbal imagery owe their meaningfulness to the fact that they stand to these *thoughts* in the unique relation of 'expressing' them. These episodes are introspectable. Indeed, it was usually believed that they could not occur without being known to occur. But this can be traced to a number of confusions, perhaps the most important of which was the idea that *thoughts* belong in the same general category as sensations, images, tickles, itches, etc. This mis-assimilation of thoughts to sensations and feelings was equally, as we saw in Sections 26 ff. above, a mis-assimilation of sensations and feelings to thoughts, and a falsification of both. The assumption that if there are thought episodes, they must be immediate experiences is common both to those who propounded the classical view and to those who reject it, saying that they 'find no such experiences.' If we purge the classical tradition of these confusions, it becomes the idea that to each of us belongs a stream of episodes, not themselves immediate experiences, to which we have privileged, but by no means either invariable or infallible, access. These episodes can occur without being 'expressed' by overt verbal behavior, though verbal behavior is—in an important sense—their natural fruition. Again, we can 'hear ourselves think,' but the verbal imagery which enables us to do this is no more the thinking itself than is the overt verbal behavior by which it is expressed and communicated to others. It is a mistake to suppose that we must be having verbal imagery—indeed, any imagery—when we 'know what we are thinking'—in short, to suppose that 'privileged access' must be construed on a perceptual or quasi-perceptual model.

Now, it is my purpose to defend such a revised classical analysis of our common-sense conception of thoughts, and in the course of

Excerpted from H. Feigl and M. Scriven, ed., *Minnesota Studies in the Philosophy of Science*, vol. 1 (University of Minnesota Press, 1956), pp. 253–329. Reprinted with permission of the publisher.

doing so I shall develop distinctions which will later contribute to a resolution, in principle, of the puzzle of *immediate experience.* But before I continue, let me hasten to add that it will turn out that the view I am about to expound could, with equal appropriateness, be represented as a modified form of the view that thoughts are *linguistic* episodes.

XII. Our Rylean Ancestors

48. But, the reader may well ask, in what sense can these episodes be 'inner' if they are not immediate experiences? and in what sense can they be 'linguistic' if they are neither overt linguistic performances, nor verbal imagery *'in foro interno'*? I am going to answer these and the other questions I have been raising by making a myth of my own, or, to give it an air of up-to-date respectability, by writing a piece of science fiction—anthropological science fiction. Imagine a stage in prehistory in which humans are limited to what I shall call a Rylean language, a language of which the fundamental descriptive vocabulary speaks of public properties of public objects located in Space and enduring through Time. Let me hasten to add that it is also Rylean in that although its basic resources are limited (how limited I shall be discussing in a moment), its total expressive power is very great. For it makes subtle use not only of the elementary logical operations of conjunction, disjunction, negation, and quantification, but especially of the subjunctive conditional. Furthermore, I shall suppose it to be characterized by the presence of the looser logical relations typical of ordinary discourse which are referred to by philosophers under the headings 'vagueness' and 'open texture.'

I am beginning my myth *in medias res* with humans who have already mastered a Rylean language, because the philosophical situation it is designed to clarify is one in which we are not puzzled by how people acquire a language for referring to public properties of public objects, but are very puzzled indeed about how we learn to speak of inner episodes and immediate experiences.

There are, I suppose, still some philosophers who are inclined to think that by allowing these mythical ancestors of ours the use *ad libitum* of subjunctive conditionals, we have, in effect, enabled them to say anything that *we* can say when we speak of *thoughts, experiences* (seeing, hearing, etc.), and *immediate experiences.* I doubt that there are many. In any case, the story I am telling is designed to show exactly *how* the idea that an intersubjective language *must* be Rylean rests on too simple a picture of the relation of intersubjective discourse to public objects.

49. The questions I am, in effect, raising are 'What resources would have to be added to the Rylean language of these talking animals in order that they might come to recognize each other and themselves as animals that *think, observe*, and have *feelings* and *sensations*, as we use these terms?' and 'How could the addition of these resources be construed as reasonable?' In the first place, the language would have to be enriched with the fundamental resources of semantical discourse—that is to say, the resources necessary for making such characteristically semantical statements as *''Rot'* means red,' and *''Der Mond ist rund'* is true if and only if the moon is round.' It is sometimes said, e.g., by Carnap 1942, that these resources can be constructed out of the vocabulary of formal logic, and that they would therefore already be contained, in principle, in our Rylean language. I have criticized this idea in another place (1963) and shall not discuss it here. In any event, a decision on this point is not essential to the argument.

Let it be granted, then, that these mythical ancestors of ours are able to characterize each other's verbal behavior in semantical terms; that, in other words, they not only can talk about each other's predictions as causes and effects, and as indicators (with greater or less reliability) of other verbal and nonverbal states of affairs, but can also say of these verbal productions that they *mean* thus and so, that they say that such and such, that they are true, false, etc. And let me emphasize, as was pointed out in Section 31 above, that to make a semantical statement about a verbal event is not a shorthand way of talking about its causes and effects, although there is a sense of 'imply' in which semantical statements about verbal productions do *imply* information about the causes and effects of these productions. Thus, when I say *''Es regnet'* means it is raining,' my statement 'implies' that the causes and effects of utterances of *'Es regnet'* beyond the Rhine parallel the causes and effects of utterances of 'It is raining' by myself and other members of the English-speaking community. And if it didn't imply this, it couldn't perform its role. But this is not to say that semantical statements are definitional shorthand for statements about the causes and effects of verbal performances.

50. With the resources of semantical discourse, the language of our fictional ancestors has acquired a dimension which gives considerably more plausibility to the claim that they are in a position to talk about *thoughts* just as we are. For characteristic of thoughts is their *intentionality, reference*, or *aboutness*, and it is clear that semantical talk about the meaning or reference of verbal expressions has the same structure as mentalistic discourse concerning what thoughts are about. It is therefore all the more tempting to suppose that the intentionality of *thoughts* can be traced to the application of semantical categories to overt verbal performances, and to suggest a modified Rylean account according to which talk about so-called 'thoughts' is shorthand for hypothetical and mongrel categorical-hypothetical statements about overt verbal and nonverbal behavior, and that talk about the *intentionality* of these 'episodes' is correspondingly reducible to semantical talk about the verbal components.

What is the alternative? Classically it has been the idea that not only are there overt verbal episodes which can be characterized in semantical terms, but, *over and above these*, there are certain inner episodes which are properly characterized by the traditional vocabulary of *intentionality*. And, of course, the classical scheme includes the idea that semantical discourse about overt verbal performances is to be analyzed in terms of talk about the intentionality of the mental episodes which are 'expressed' by these overt performances. My immediate problem is to see if I can reconcile the classical idea of thoughts as inner episodes which are neither overt behavior nor verbal imagery and which are properly referred to in terms of the vocabulary of intentionality, with the idea that the categories of intentionality are, at bottom, semantical categories pertaining to overt verbal performances.[1]

XIII. Theories and Models

51. But what might these episodes be? And, in terms of our science fiction, how might our ancestors have come to recognize their existence? The answer to these questions is surprisingly straightforward, once the logical space of our discussion is enlarged to include a distinction, central to the philosophy of science, between the language of *theory* and the language of *observation*. Although this distinction is a familiar one, I shall take a few paragraphs to highlight those aspects of the distinction which are of greatest relevance to our problem.

Informally, to construct a theory is, in its most developed or sophisticated form, to postulate a domain of entities which behave in certain ways set down by the fundamental principles of the theory, and to correlate—perhaps, in a certain sense to identify—complexes of these theoretical entities with certain nontheoretical objects or situations; that is to say, with objects or situations which are either matters of observable fact or, in principle at least, describable in observational terms. This 'correlation' or 'identification' of theoretical with observational states of affairs is a tentative one 'until further notice,' and amounts, so to speak, to erecting temporary bridges which permit the passage from sentences in observational discourse to sentences in the theory, and vice versa. Thus, for example, in the kinetic theory of gases, empirical statements of the form 'Gas g at such and such a place and time has such and such a volume, pressure, and temperature' are correlated with theoretical statements specifying certain statistical measures of populations of molecules. These temporary bridges are so set up that inductively established laws pertaining to gases, formulated in the language of observable fact, are correlated with derived propositions or theorems in the language of the theory, and that no proposition in the theory is correlated with a falsified empirical generalization. Thus, a good theory (at least of the type we are considering) 'explains' established empirical laws by deriving theoretical counterparts of these laws from a small set of postulates relating to unobserved entities.

These remarks, of course, barely scratch the surface of the problem of the status of theories in scientific discourse. And no sooner have I made them, than I must hasten to qualify them—almost beyond recognition. For while this by now classical account of the nature of theories (one of the earlier formulations of which is due to Norman Campbell 1920, and which is to be bound more recently in the writings of Carnap 1953; Reichenbach 1928, 1938; Hempel 1952; and Braithwaite 1953) does throw light on the logical status of theories, it emphasizes certain features at the expense of others. By speaking of the construction of a theory as the elaboration of a postulate system which is tentatively correlated with observational discourse, it gives a highly artificial and unrealistic picture of what scientists have actually done in the process of constructing

theories. I don't wish to deny that logically sophisticated scientists today *might* and perhaps, on occasion, do proceed in true logistical style. I do, however, wish to emphasize two points:

(1) The first is that the fundamental assumptions of a theory are usually developed not by constructing uninterpreted calculi which might correlate in the desired manner with observational discourse, but rather by attempting to find a *model*, i.e. to describe a domain of familiar objects behaving in familiar ways such that we can see how the phenomena to be explained would arise if they consisted of this sort of thing. The essential thing about a model is that it is accompanied, so to speak, by a commentary which *qualifies* or *limits*—but not precisely nor in all respects—the analogy between the familiar objects and the entities which are being introduced by the theory. It is the descriptions of the fundamental ways in which the objects in the model domain, thus qualified, behave, which, transferred to the theoretical entities, correspond to the postulates of the logistical picture of theory construction.

(2) But even more important for our purposes is the fact that the logistical picture of theory construction obscures the most important thing of all, namely that the process of devising 'theoretical' explanations of observable phenomena did not spring full-blown from the head of modern science. In particular, it obscures the fact that not all common-sense inductive inferences are of the form

> All observed A's have been B, *therefore* (*probably*) all A's are B.

or its statistical counterparts, and leads one mistakenly to suppose that so-called 'hypotheticdeductive' explanation is limited to the sophisticated stages of science. The truth of the matter, as I shall shortly be illustrating, is that science is continuous with common sense, and the ways in which the scientist seeks to explain empirical phenomena are refinements of the ways in which plain men, however crudely and schematically, have attempted to understand their environment and their fellow men since the dawn of intelligence. It is this point which I wish to stress at the present time, for I am going to argue that the distinction between theoretical and observational discourse is involved in the logic of concepts pertaining to inner episodes. I say 'involved in' for it would be paradoxical and, indeed, incorrect, to say that these concepts *are* theoretical concepts.

52. Now I think it fair to say that some light has already been thrown on the expression 'inner episodes'; for while it would indeed be a category mistake to suppose that the inflammability of a piece of wood is, so to speak, a hidden burning which becomes overt or manifest when the wood is placed on the fire, not all the unobservable episodes we suppose to go on in the world are the offspring of category mistakes. Clearly it is by no means an illegitimate use of 'in'—though it is a use which has its own logical grammar—to say, for example, that 'in' the air around us there are innumerable molecules which, in spite of the observable stodginess of the air, are participating in a veritable turmoil of episodes. Clearly, the sense in which these episodes are 'in' the air is to be explicated in terms of the sense in which the air 'is' a population of molecules, and this, in turn, in terms of the logic of the relation between theoretical and observational discourse.

I shall have more to say on this topic in a moment. In the meantime, let us return to our mythical ancestors. It will not surprise my readers to learn that the second stage in the enrichment of their Rylean language is the addition of theoretical discourse. Thus we may suppose these language-using animals to elaborate, without methodological sophistication, crude, sketchy, and vague theories to explain why things which are similar in their observable properties differ in their causal properties, and things which are similar in their causal properties differ in their observable properties.

XIV. Methodological versus Philosophical Behaviorism

53. But we are approaching the time for the central episode in our myth. I want you to suppose that in this Neo-Rylean culture there now appears a genius—let us call him Jones—who is an unsung forerunner of the movement in psychology, once revolutionary, now commonplace, known as Behaviorism. Let me emphasize that what I have in mind is Behaviorism as a methodological thesis, which I shall be concerned to formulate. For the central and guiding theme in the historical complex known by this term has been a certain conception, or family of conceptions, of how to go about building a science of psychology.

Philosophers have sometimes supposed that Behaviorists are, as such, committed to the idea that our ordinary mentalistic concepts

are *analyzable* in terms of overt behavior. But although behaviorism has often been characterized by a certain metaphysical bias, it is not a thesis about the *analysis* of *existing* psychological concepts, but one which concerns the construction of new concepts. As a methodological thesis, it involves no commitment whatever concerning the logical analysis of common-sense mentalistic discourse, nor does it involve a denial that each of us has a privileged access to our state of mind, nor that these states of mind can properly be described in terms of such common-sense concepts as believing, wondering, doubting, intending, wishing, inferring, etc. If we permit ourselves to speak of this privileged access to our states of mind as 'introspection,' avoiding the implication that there is a 'means' whereby we 'see' what is going on 'inside,' as we see external circumstances by the eye, then we can say that Behaviorism, as I shall use the term, does not deny that there is such a thing as introspection, nor that it is, on some topics, at least, quite reliable. The essential point about 'introspection' from the standpoint of Behaviorism is that *we introspect in terms of common sense mentalistic concepts.* And while the Behaviorist admits, as anyone must, that much knowledge is embodied in common-sense mentalistic discourse, and that still more can be gained in the future by formulating and testing hypotheses in terms of them, and while he admits that it is perfectly legitimate to call such a psychology 'scientific,' he proposes, for his own part, to make no more than a heuristic use of mentalistic discourse, and to construct his concepts 'from scratch' in the course of developing his own scientific account of the observable behavior of human organisms.

54. But while it is quite clear that scientific Behaviorism is *not* the thesis that common-sense psychological concepts are *analyzable* into concepts pertaining to overt behavior—a thesis which has been maintained by some philosophers and which may be called 'analytical' or 'philosophical' Behaviorism—it is often thought that Behaviorism is committed to the idea that the concepts of a behavioristic psychology must be so analyzable, or, to put things right side up, that properly introduced behavioristic concepts must be built by explicit definition—in the broadest sense—from a basic vocabulary pertaining to overt behavior. The Behaviorist would thus be saying 'Whether or not the mentalistic concepts of everyday life are definable in terms of overt behavior, I shall ensure that this is true of the concepts that I shall employ.' And it must be confessed that many behavioristically oriented psychologists have believed themselves committed to this austere program of concept formation.

Now I think it reasonable to say that, *thus conceived*, the behavioristic program would be unduly restrictive. Certainly, nothing in the nature of sound scientific procedure requires this self-denial. Physics, the methodological sophistication of which has so impressed—indeed, overly impressed—the other sciences, does not lay down a corresponding restriction on its concepts, nor has chemistry been built in terms of concepts explicitly definable in terms of the observable properties and behavior of chemical substances. The point I am making should now be clear. The behavioristic requirement that all concepts should be *introduced* in terms of a basic vocabulary pertaining to overt behavior is compatible with the idea that some behavioristic concepts are to be introduced as *theoretical* concepts.

55. It is essential to note that the theoretical terms of a behavioristic psychology are not only *not* defined in terms of overt behavior, they are also *not* defined in terms of nerves, synapses, neural impulses, etc., etc. A behavioristic theory of behavior is not, as such, a physiological explanation of behavior. The ability of a framework of theoretical concepts and propositions successfully to explain behavioral phenomena is logically independent of the identification of these theoretical concepts with concepts of neurophysiology. What *is* true—and this is a logical point—is that each special science dealing with some aspect of the human organism operates within the frame of a certain regulative ideal, the ideal of a coherent system in which the achievements of each have an intelligible place. Thus, it is part of the Behaviorist's business to keep an eye on the total picture of the human organism which is beginning to emerge. And if the tendency to premature identification is held in check, there may be considerable heuristic value in speculative attempts at integration; though, until recently, at least, neurophysiological speculations in behavior theory have not been particularly fruitful. And while it is, I suppose, noncontroversial that when the total scientific picture of man and his behavior is in, it will involve *some* identification of concepts in behavior theory with concepts pertaining to the functioning of anatomical structures, it should not be assumed that behavior theory is committed *ab initio* to a physiological identification

of *all* its concepts,—that its concepts are, so to speak, physiological from the start.

We have, in effect, been distinguishing between two dimensions of the logic (or 'methodologic') of theoretical terms: (a) their role in explaining the selected phenomena of which the theory is the theory; (b) their role as candidates for integration in what we have called the 'total picture.' These roles are equally part of the logic, and hence the 'meaning,' of theoretical terms. Thus, at any one time the terms in a theory will carry with them as part of their logical force that which it is reasonable to envisage—whether schematically or determinately—as the manner of their integration. However, for the purposes of my argument, it will be useful to refer to these two roles as though it were a matter of a distinction between what I shall call *pure theoretical concepts*, and hypotheses concerning the relation of these concepts to concepts in other specialties. What we *can* say is that the less a scientist is in a position to conjecture about the way in which a certain theory can be expected to integrate with other specialities, the more the concepts of his theory approximate to the status of pure theoretical concepts. To illustrate: We can imagine that Chemistry developed a sophisticated and successful theory to explain chemical phenomena before either electrical or magnetic phenomena were noticed; and that chemists developed as pure theoretical concepts, certain concepts which it later became reasonable to identify with concepts belonging to the framework of electromagnetic theory.

XV. The Logic of Private Episodes: Thoughts

56. With these all too sketchy remarks on Methodological Behaviorism under our belts, let us return once again to our fictional ancestors. We are now in a position to characterize the original Rylean language in which they described themselves and their fellows as not only a *behavioristic* language, but a behavioristic language which is restricted to the *non-theoretical* vocabulary of a behavioristic psychology. Suppose, now, that in the attempt to account for the fact that his fellow men behave intelligently not only when their conduct is threaded on a string of overt verbal episodes—that is to say, as we would put it, when they 'think out loud'—but also when no detectable verbal output is present, Jones develops a *theory* according to which overt utterances are but the culmination of a process which begins with certain inner episodes. *And let us suppose that his model for these episodes* which initiate the events which culminate in overt verbal behavior *is that of overt verbal behavior itself. In other words, using the language of the model, the theory is to the effect that overt verbal behavior is the culmination of a process which begins with 'inner speech.'*

It is essential to bear in mind that what Jones means by 'inner speech' is not to be confused with *verbal imagery.* As a matter of fact, Jones, like his fellows, does not as yet even have the concept of an image.

It is easy to see the general lines a Jonesean theory will take. According to it the true cause of intelligent nonhabitual behavior is 'inner speech.' Thus, even when a hungry person overtly says 'Here is an edible object' and proceeds to eat it, the true—theoretical—cause of his eating, given his hunger, is not the overt utterance, but the 'inner utterance of this sentence.'

57. The first thing to note about the Jonesean theory is that, as built on the model of speech episodes, *it carries over to these inner episodes the applicability of semantical categories.* Thus, just as Jones has, like his fellows, been speaking of overt utterances as *meaning* this or that, or being *about* this or that, so he now speaks of these inner episodes as *meaning* this or that, or being *about* this or that.

The second point to remember is that although Jones' theory involves a *model*, it is not identical with it. Like all theories formulated in terms of a model, it also includes a *commentary* on the model; a commentary which places more or less sharply drawn restrictions on the analogy between the theoretical entities and the entities of the model. Thus, while his theory talks of 'inner speech,' the commentary hastens to add that, of course, the episodes in question are not the wagging of a hidden tongue, nor are any sounds produced by this 'inner speech.'

58. The general drift of my story should now be clear. I shall therefore proceed to make the essential points quite briefly:

(1) What we must suppose Jones to have developed is the germ of a theory which permits many different developments. We must not pin it down to any of the more sophisticated forms it takes in the hands of classical philosophers. Thus, the theory need not be given a Socratic or Cartesian form, according to which this 'inner speech' is a function of a separate substance;

though primitive peoples may have had good reason to suppose that humans consist of two separate things.

(2) Let us suppose Jones to have called these discursive entities *thoughts*. We can admit at once that the framework of thoughts he has introduced is a framework of 'unobserved,' 'nonempirical' 'inner' episodes. For we can point out immediately that in these respects they are no worse off than the particles and episodes of physical theory. For these episodes are 'in' language-using animals as molecular impacts are 'in' gases, not as 'ghosts' are in 'machines.' They are 'nonempirical' in the simple sense that they are *theoretical*—not definable in observational terms. Nor does the fact that they are, *as introduced*, unobserved entities imply that Jones could not have good reason for supposing them to exist. Their 'purity' is not a *metaphysical* purity, but, so to speak, a *methodological* purity. As we have seen, the fact that they are not introduced as physiological entities does not preclude the possibility that at a later methodological stage, they may, so to speak, 'turn out' to be such. Thus, there are many who would say that it is already reasonable to suppose that these *thoughts* are to be 'identified' with complex events in the cerebral cortex functioning along the lines of a calculating machine. Jones, of course, has no such idea.

(3) Although the theory postulates that overt discourse is the culmination of a process which begins with 'inner discourse,' this should not be taken to mean that overt discourse stands to 'inner discourse' *as voluntary movements stand to intentions and motives.* True, overt linguistic events *can* be produced as means to ends. But serious errors creep into the interpretation of both language and thought if one interprets the idea that overt linguistic episodes *express* thoughts, on the model of the use of an instrument. Thus, it should be noted that Jones' theory, as I have sketched it, is perfectly compatible with the idea that the ability to have thoughts is acquired in the process of acquiring overt speech and that only after overt speech is well established, can 'inner speech' occur without its overt culmination.

(4) Although the occurrence of overt speech episodes which are characterizable in semantical terms is explained by the theory in terms of *thoughts* which are *also* characterized in semantical terms, this does not mean that the idea that overt speech 'has meaning' is being *analyzed* in terms of the intentionality of thoughts. It must not be forgotten that *the semantical characterization of overt verbal episodes is the primary use of semantical terms, and that overt linguistic events as semantically characterized are the model for the inner episodes introduced by the theory.*

(5) One final point before we come to the dénouement of the first episode in the saga of Jones. It cannot be emphasized too much that although these theoretical discursive episodes or *thoughts* are introduced as *inner* episodes—which is merely to repeat that they are introduced as *theoretical* episodes—they are *not* introduced as *immediate experiences.* Let me remind the reader that Jones, like his Neo-Rylean contemporaries, does not as yet have this concept. And even when he, and they, acquire it, by a process which will be the second episode in my myth, it will only be the philosophers among them who will suppose that the inner episodes introduced for one theoretical purpose—thoughts—must be a subset of immediate experiences, inner episodes introduced for another theoretical purpose.

59. Here, then, is the dénouement. I have suggested a number of times that although it would be most misleading to say that concepts pertaining to thinking are theoretical concepts, yet their status might be illuminated by means of the contrast between theoretical and nontheoretical discourse. We are now in a position to see exactly why this is so. For once our fictitious ancestor, Jones, has developed the theory that overt verbal behavior is the expression of thoughts, and taught his compatriots to make use of the theory in interpreting each other's behavior, it is but a short step to the use of this language in self-description. Thus, when Tom, watching Dick, has behavioral evidence which warrants the use of the sentence (in the language of the theory) 'Dick is thinking 'p' ' (or 'Dick is thinking that p'), Dick, using the same behavioral evidence, can say, in the language of the theory, 'I am thinking 'p' ' (or 'I am thinking that p.') And it now turns out—need it have?—that Dick can be trained to give reasonably reliable self-descriptions, using the language of the theory, without having to observe his overt behavior. Jones brings this about, roughly, by applauding utterances by Dick of 'I am thinking that p' when the behavioral evidence strongly supports the theoretical statement 'Dick is thinking that p'; and by frowning on utterances of 'I am thinking that p,' when the evidence does not support this theoretical statement. Our ancestors begin to speak of the privileged access each of us has to his own thoughts.

What began as a language with a purely theoretical use has gained a reporting role.

As I see it, this story helps us understand that concepts pertaining to such inner episodes as thoughts are primarily and essentially *intersubjective*, as intersubjective as the concept of a positron, and that the reporting role of these concepts—the fact that each of us has a privileged access to his thoughts—constitutes a dimension of the use of these concepts which is *built on* and *presupposes* this intersubjective status. My myth has shown that the fact that language is essentially an *intersubjective* achievement, and is learned in intersubjective contexts—a fact rightly stressed in modern psychologies of language, thus by B. F. Skinner 1945, and by certain philosophers, e.g. Carnap 1933, Wittgenstein 1953—is compatible with the 'privacy' of 'inner episodes.' It also makes clear that this privacy is not an 'absolute privacy.' For if it recognizes that these concepts have a reporting use in which one is not drawing inferences from behavioral evidence, it nevertheless insists that the fact that overt behavior *is* evidence for these episodes *is built into the very logic of these concepts*, just as the fact that the observable behavior of gases is evidence for molecular episodes is built into the very logic of molecule talk. . . .

REFERENCES

Braithwaite, R. B. *Scientific Explanation* (Cambridge, UK: Cambridge University Press, 1953).

Campbell, Norman. *Physics: The Elements* (Cambridge, UK: Cambridge University Press, 1920).

Carnap, Rudolf. *Introduction to Semantics* (Chicago: University of Chicago Press, 1942).

_____"*Psychologie in Physikalischer Sprache*," *Erkenntnis* 3 (1933): pp. 107–42.

_____ "The Interpretation of Physics," in *Readings in the Philosophy of Science*, H. Feigl and M. Brodbeck, eds., (New York: Appleton-Century-Crofts, 1953), pp. 309–18. This selection consists of pp. 59–69 of his *Foundations of Logic and Mathematics* (Chicago: University of Chicago Press, 1939).

Hempel, C. G. *Fundamentals of Concept Formation in Empirical Science* (Chicago: University of Chicago Press), 1952.

Reichenbach, H. *Philosophie der Raum-Zeit-Lehre* (Berlin: de Gruyter, 1928).

_____ *Experience and Prediction* (Chicago: University of Chicago Press, 1938).

Sellars, Wilfrid. "Mind, Meaning and Behavior," *Philosophical Studies* 3 (1952): pp. 83–94.

_____ "A Semantical Solution of the Mind-Body Problem," *Methodos* 5 (1953): pp. 45–84.

_____ "Empiricism and Abstract Entities," in *The Philosophy of Rudolf Carnap*, Paul A. Schlipp, ed., (Evanston, IL: Library of Living Philosophers, 1963), pp. 431–68.

Skinner, B. F., "The Operational Analysis of Psychological Terms," *Psychological Review* 52 (1945): pp. 270–77. Reprinted in *Readings in the Philosophy of Science*, H. Feigl and M. Brodbeck, eds., (New York: Appleton-Century-Crofts, 1953), pp. 585–94.

Wittgenstein, Ludwig. *Philosophical Investigations* (London: Macmillan, 1953).

NOTE

This paper was first presented as the University of London Special Lectures on Philosophy for 1955–56, delivered on March 1, 8, and 15, 1956, under the title "The Myth of the Given: Three Lectures on Empiricism and the Philosophy of Mind."

1. An earlier attempt along these lines is to be found in my 1952 and 1953.

Propositional Attitudes

Jerry A. Fodor

Some philosophers (Dewey, for example, and maybe Austin) hold that philosophy is what you do to a problem until it's clear enough to solve it by doing science. Others (Ryle, for example, and maybe Wittgenstein) hold that if a philosophical problem succumbs to empirical methods, that shows it wasn't *really* philosophical to begin with. Either way, the facts seem clear enough: questions first mooted by philosophers are sometimes co-opted by people who do experiments. This seems to be happening now to the question: 'what are propositional attitudes?' and cognitive psychology is the science of note.

One way to elucidate this situation is to examine theories that cognitive psychologists endorse, with an eye to explicating the account of propositional attitudes that the theories presuppose. That was my strategy in Fodor 1975. In this paper, however, I'll take another tack. I want to outline a number of a priori conditions which, on my view, a theory of propositional attitudes (PAs) ought to meet. I'll argue that, considered together, these conditions pretty clearly demand a treatment of PAs as relations between organisms and internal representations; precisely the view that the psychologists have independently arrived at. I'll thus be arguing that we have good reasons to endorse the psychologists' theory even aside from the empirical exigencies that drove them to it. I take it that this convergence between what's plausible a priori and what's demanded ex post facto is itself a reason for believing that the theory is probably true.

Three preliminary remarks: first, I'm not taking 'a priori' all that seriously. Some of the points I'll be making are, I suppose, strictly conceptual, but others are merely self-evident. What I've got is a set of glaring facts about propositional attitudes. I don't doubt that we might rationally adopt an account of the attitudes which contravenes some, or maybe even all of them. But the independent evidence for such an account would have to be extremely persuasive or I, for one, would get the jitters. Second, practically everything I'll say about the attitudes has been said previously in the philosophical literature. All I've done is bring the stuff together. I do think, however, that the various constraints that I'll discuss illuminate each other; it is only when one attempts to satisfy them all at once that one sees how univocal their demands are. Finally, though I intend what I say to apply, mutatis mutandis, to PAs at large, I shall run the discussion pretty much exclusively on beliefs and wants. These seem to be the root cases for a systematic cognitive psychology; thus learning and perception are presumably to be treated as varieties of the fixation of belief, and the theory of action is presumably continuous with the theory of utility.[1]

Here, then, are my conditions, with comments.

I. Propositional attitudes should be analyzed as relations. In particular, the verb in a sentence like 'John believes it's raining' expresses a relation between John and something else, and a token of that sentence is true if John stands in the belief-relation to that thing.[2] Equivalently, for these purposes, 'it's raining' is a term in 'John believes it's raining.'[3] I have three arguments for imposing condition I, all of them inconclusive.

I-a) It's intuitively plausible. 'Believes' looks like a two-place relation, and it would be nice if our theory of belief permitted us to save the appearances.

No doubt, appearances sometimes deceive. The ''s' in 'Mary's sake' looks like expressing a relation (of possession) between Mary and a sake; but it doesn't, or so we're told. In fact, 'Mary's sake' doesn't look *very* relational, since *x's sake* would surely qualify as an idiom even if we had no ontological scruples to placate. There's something syntactically wrong with: ⌜Mary's sake is *Fer* than Bill's⌝, 'Mary has a (little) sake,' etc. For that matter, there's something syntactically wrong with 'a sake' *tout court.* Yet, we'd expect all such expressions to be well-formed if 'Mary's sake' contained

From *The Monist* 61 (1978): pp. 573–91.

a true possessive. 'Mary's sake' doesn't bear comparison with 'Mary's lamb.'

Still, there are some cases of *non*-idiomatic expressions which appear to be relational, but which, upon reflection, maybe aren't. 'Mary's voice' goes through the transformations even if 'Mary's sake' does not (Dennett 1969). Yet there aren't, perhaps, such *things* as voices; and, if there aren't, 'Mary's voice' can't refer in virtue of a relation between Mary and one of them.[4] I think it is fair to view the 'surface' grammar as ontologically misleading in *these* cases, but only because we know how to translate into more parsimonious forms. 'Mary has a good voice (bad voice; little voice; better voice than Bill's)' goes over, pretty much without residue, into 'Mary sings well (badly, weakly, less well than Bill).' If, however, we were *un*able to provide (or, anyhow, to envision providing) the relevant translations, what right would we have to view such expressions as ontologically promiscuous? 'Bill believes it's raining' is not an idiom, and there is, so far as anybody knows, no way of translating sentences nominally about beliefs into sentences of reduced ontological load. (Behaviorists used to think such translations might be forthcoming, but they were wrong.) We must, then, either take the apparent ontological commitments seriously or admit to playing fast and loose.

I-b) Existential Generalization applies to the syntactic objects of verbs of propositional attitude; from 'John believes it's raining' we can infer 'John believes something' and 'there is something that John believes' (viz., that it's raining). *EG* may not be *criterial* for ontological commitment, but it is surely a straw in the wind.[5]

I-c) The only known alternative to the view that verbs of propositional attitude express relations is that they are (semantically) 'fused' with their objects, and that view would seem to be hopeless.[6]

The fusion story is the proposal that sentences like 'John believes it's raining' ought really to be spelled 'John believes-it's-raining'; that the logical form of such sentences acknowledges a referring expression ('John') and a one-place predicate with no internal structure ('believes-it's-raining'). 'John believes it's raining' is thus an atomic sentence, similar *au fond* to 'John is purple.'

Talk about counter-intuitive! Moreover:

1. There are infinitely many (semantically distinct) sentences of the form *a believes complement.* If all such sentences are atomic, how is English learned? (Davidson 1965).
2. Different propositional attitudes are often 'focused' on the same content; for example, one can both fear and believe that it will rain on Tuesday. But, on the fusion view, 'John fears that it will rain on Tuesday' has nothing in common with 'John believes that it will rain on Tuesday' save only the reference to John. In particular, it's an *accident* that the form of words 'it will rain on Tuesday' occurs in both.
3. Similarly, different beliefs can be related in such ways as the following: John thinks Sam is nice; Mary thinks Sam is nasty. Under ordinary English representation these beliefs overlap at the 'Sam' position, so the notation sustains the intuition that John and Mary disagree about Sam. But, if the fusion view is correct, 'John thinks Sam is nice' and 'Mary thinks Sam is nasty' have no more in common at the level of canonical notation than, say, 'John eats' and 'Mary swims.' Talk about imperspicuous! In respect of saving the intuitions, the recommended reconstruction does *worse* than the undisciplined orthography that we started with.[7] (For that matter, there's nothing in ⌜believes-that-S⌝ to suggest that it's about believing. Here too ⌜believes that S⌝ does much better.)
4. It could hardly be an accident that the declarative sentences of English constitute the (syntactic) objects of verbs like 'believe.' Whereas, on the fusion view it's *precisely* an accident; the complement of 'believes' in 'John believes it's raining' bears no more relation to the sentence 'It's raining' than, say, the word 'dog' bears to the first syllable of 'dogmatic.'
5. On the fusion view, it's a sheer accident that if 'John believes it's raining' is true, then what John believes is true if 'it's raining' is true. But this, surely, is one accident too many. Surely the identity between the truth conditions on John's belief when he believes Fa, and those on the corresponding sentence ⌜a is F⌝ must be what connects the theory of sentence interpretation with the theory of PAs (and what explains our using 'it's raining,' and not some other form of words, to specify *which* belief John has when he believes it's raining).

It's the mark of a bad theory that it makes the data look fortuitous. I conclude that the fusion

story is not to be taken very seriously; that neither the philosophy of language nor the philosophy of mind is advanced just by proliferating hyphens. But the fusion story is (de facto) the only alternative to the view that 'believe' expresses a relation. Hence, first blush, we had better assume that 'believe' *does* express a relation and try to find an account of propositional attitudes which comports with that assumption.

II. A theory of PAs should explain the parallelism between verbs of PA and verbs of saying. ("Vendler's Condition").

Rather generally, the things we can be said to *believe* (want, hope, regret, etc.) are the very things that we can be said to *say* (assert, state, etc.). So, John can either believe or assert that it's about to blow; he can either hope that or inquire whether somebody has reefed the main; he can either doubt or demand that the crew should douse the Genny. Moreover, as Vendler 1972 has shown, there are interesting consequences of classifying verbs of PA (on the one hand) and verbs of saying (on the other) by reference to the syntax of their object complements. It turns out that the taxonomies thus engendered are isomorphic down to surprisingly fine levels of grain. Now, of course, this *could* be just an accident, as could the semantic and syntactic parallelisms between the complements of verbs of PA and free-standing declaratives (see above). Certainly, it's a substantial inference from the syntactic similarities that Vendler observes to the conclusion he draws: that the object of assertion is identical with the object of belief. Suffice it for now to make the less ambitious point: we should prefer a theory which explains the facts to one which merely shrugs its shoulders; viz. a theory which satisfies Vendler's condition to a theory which does not.

III. A theory of propositional attitudes should account for their opacity ("Frege's Condition").

Thus far, I have stressed logico-syntactic analogies between the complements of belief clauses and the corresponding free-standing declaratives. However, it has been customary in the philosophical literature since Frege to stress one of their striking *dis*analogies: the former are, in general, opaque to inferential operations to which the latter are, in general, transparent. Since this aspect of the behavior of sentences that ascribe propositional attitudes has so dominated the philosophical discussion, I shall make the point quite briefly here. Sentences containing verbs of PA are not, normally, truth functions of their complements. Moreover, contexts subordinated to verbs of PA are normally themselves non-truth functional, and *EG* and substitution of identicals may apply at syntactic positions in a free-standing declarative while failing at syntactically comparable positions in belief sentences. A theory of PAs should explain why all this is so.

It should be acknowledged that, however gross the inadequacies of the fusion view, it does at least provide an account of propositional attitudes which meets Frege's condition. If *S* doesn't so much as occur in ⌜John believes S⌝ it's hardly surprising that the one should fail to be a truth function of the other; similarly, if 'Mary' doesn't occur in 'Bill believes that John bit Mary,' it's hardly surprising that the sentence doesn't behave the way it would if 'Mary' occurred referentially. The methodological moral is perhaps that Frege's condition underconstrains a theory of PAs; ideally, an acceptable account of opacity should follow from a theory that is independently plausible.

IV. The objects of propositional attitudes have logical form ("Aristotle's Condition").

Mental states (including, especially, token havings of propositional attitudes) interact causally. Such interactions constitute the mental processes which eventuate (inter alia) in the behaviors of organisms. Now, it is crucial to the whole program of explaining behavior by reference to mental states that the propositional attitudes belonging to these chains are typically *non*-arbitrarily related in respect of their content (taking the 'content' of a propositional attitude, informally, to be whatever it is that the complement of the corresponding PA-ascribing sentence expresses).

This is not an a priori claim, though perhaps it is a transcendental one. For, though one can imagine the occurrence of causal chains of mental states which are not otherwise related (as, e.g., a thought that two is a prime number, causing a desire for tea, causing an intention to recite the alphabet backwards, causing an expectation of rain) and though such sequences doubtless actually occur (in dreams, say, and in madness) still if *all* our mental life were like this, it's hard to see what point ascriptions of contents to mental states would have. Even phenomenology presupposes some correspondence between the content of our beliefs and the content of our beliefs about our beliefs; else

there would be no coherent introspections for phenomenologists to report.

The paradigm situation—the grist for the cognitivist's mill—is the one where propositional attitudes interact causally and do so *in virtue of* their content. And the paradigm of this paradigm is the practical syllogism. Since it is part of my point that the details matter not at all, I shall take liberties with Aristotle's text.

John believes that it will rain if he washes his car. John wants it to rain. So John acts in a manner intended to be a car-washing.

I take it that this might be a true, if informal, etiology of John's 'car-washing behavior'; the car washing is an effect of the intention to car-wash, and the intention to car-wash is an effect of the causal interaction between John's beliefs and his utilities. Moreover, the etiological account might be counterfactual-supporting in at least the following sense: John wouldn't have car-washed had the content of his beliefs, utilities and intentions been other than they were. Or, if he did, he would have done so unintentionally, or for different reasons, or with other ends in view. To say that John's mental states interact causally *in virtue of* their content is, in part, to say that such counterfactuals hold.

If there are true, contingent counterfactuals which relate mental state *tokens* in virtue of their contents, that is presumably because there are true, contingent generalizations which relate mental state *types* in virtue of their contents. So, still following Aristotle at a distance, we can schematize etiologies like the one above to get the underlying generalization: if x believes that A is an action x can perform; and if x believes that a performance of A is sufficient to bring it about that Q; and if x wants it to be the case that Q; then x acts in a fashion intended be a performance of A.

I am not, for present purposes, interested in whether this is a plausible decision theory; still less in whether it is the decision theory that Aristotle thought plausible. What interests me here is rather: (a) that any decision theory we can now contemplate will surely look rather like this one in that (b) it will entail generalizations about the causal relations among content-related beliefs, utilities and intentions; and (c) such generalizations will be specified by reference to the form of the propositional attitudes which instantiate them. (This remains true even if, as some philosophers suppose, an adequate decision theory is irremediably in need of ceteris paribus clauses to flesh out its generalizations. See, for example, Grice 1975.) So, in particular, we can't state the theory-relevant generalization that is instantiated by the relations among John's mental states unless we allow reference to beliefs of the form *if X then Y*, desires of the form *that Y;* intentions of the form *that X should come about;* and so forth. Viewed one way (material mode) the recurrent schematic letters require identities of content among propositional attitudes. Viewed the other way (linguistically) they require formal identities among the complements of the PA-ascribing sentence which instantiate the generalizations of the theory that explains John's behavior. Either way, the form of the generalization determines how the theory relates to the events that it subsumes. There is nothing remarkable about this, of course, except that form is here being ascribed *inside* the scope of verbs of PA.

To summarize: our common-sense psychological generalizations relate mental states in virtue of their content, and canonical representation does what it can to reconstruct such content relations as relations of form. 'Aristotle's condition' requires that our theory of propositional attitudes should rationalize this process by construing verbs of PA in a way that permits reference to the form of their objects. To do this is to legitimize the presuppositions of commonsense psychology and, for that matter, of real (viz. cognitive) psychology as well. (See Fodor, op. cit.)

In fact, we can state (and satisfy) Aristotle's condition in a still stronger version. Let anything be a *belief sentence* if it is of the form *a believes that S.* Define the *correspondent* of such a sentence as the formula which consists of *S* standing alone (i.e. the sentence #S#).[8] We remarked above that there is the following relation between the truth conditions on the belief that a belief sentence ascribes and the truth conditions on the correspondent of the belief sentence: the belief is true if the correspondent is. This is, presumably, at least part of what is involved in viewing the correspondent of a belief sentence as *expressing* the ascribed belief.

It should not, therefore, be surprising to find that our intuitions about the form of the belief ascribed by a given belief sentence are determined by the logical form of its correspondent. So, intuitively, John's belief that Mary and Bill are leaving is a conjunctive belief (cf. the logical form of 'Mary and Bill are leaving'); John's belief that Alfred is a white swan is a singulary belief (cf. the logical form of 'Alfred is a white swan'); and so on. It is, of course, essential

that we understand 'belief' *opaquely* in such examples; otherwise, the belief that P will have the logical form of any sentence equivalent to P. But this is as it should be: it is in virtue of its *opaque* content that John's belief that P plays its systematic role in John's mental life: e.g., in the determination of his actions and in the causation of his other mental states. Hence it is the opaque construal that operates in such patterns of explanation as the practical syllogism and its spiritual heirs.

We are now in position to state Aristotle's condition in its strongest (and final) version. A theory of propositional attitudes should legitimize the ascription of form to objects of propositional attitudes. In particular, it should explain why the form of a belief is identical to the logical form of the correspondent of a sentence which (opaquely) ascribes that belief.[9]

I digress: One may feel inclined to argue that the satisfaction of Aristotle's condition is incompatible with the satisfaction of Frege's condition; that the opacity of belief sentences shows the futility of assigning logical form to their objects. The argument might go as follows. Sentence have logical form in virtue of their behavior under logical transformations; the logical form of a sentence is that aspect of its structure in virtue of which it provides a domain for such transformations. But Frege shows us that the objects of verbs of propositional attitude are inferentially inert. Hence, it's a sort of charade to speak of the logical form of the objects of PAs; what's the force of saying that a sentence has the form *P & Q* if one must also say that simplification of conjunction does not apply?

Perhaps some such argument supplies the motive force of fusion theories. It is, in any event, misled. In particular, it muddles the distinction between what's entailed by what's believed, and what's entailed by believing what's believed. Less cryptically: if John believes that P & Q, then what John believes entails that P and what John believes entails that Q. This is surely incontestable; P & Q is what John believes, and P & Q entails P, Q. Full stop. It would thus be highly ill-advised to put Frege's condition as 'P & Q is semantically inert when embedded to the context ⌜John believes . . .⌝'; for this makes it sound as though P & Q sometimes doesn't entail P: viz. when it's in the scope of 'believes.' (A parallel bad argument: P & Q sometimes doesn't entail P, viz. when it's in the scope of the operator 'not) What falls under Frege's condition, then, is not the sentence that expresses what John believes (viz. P & Q) but the sentence that expresses John's believing what he believes (viz. the sentence ⌜John believes that P & Q⌝). Note that the inertia of this latter sentence isn't an exception to simplification of conjunction since simplification of conjunction isn't defined for sentences of the form *a believes that P & Q;* only for sentences of the form *P & Q*.

'Still,' one might say, 'if the form of words ⌜P & Q⌝ is logically inert when embedded to the form of words ⌜John believes . . .⌝, what's the *point* of talking about the logical form of the complement of belief sentences?' This isn't an argument, of course, but it's a fair question. Answers: (a) because we may want to satisfy Aristotle's condition (e.g., in order to be in a position to state the practical syllogism); (b) because we may want to compare beliefs in respect of their form (John's belief that [x] Fx → Gx is a generalization of Mary's belief that a is F and G; Sam's belief that P is incompatible with Bill's belief that not-P; etc.); (c) because we may wish to speak of the consequences of a belief, even while cheerfully admitting that the consequences of a belief may not themselves be objects of belief (viz. believed in). Indeed, we need the notion of the consequences of a belief if only in order to say that belief isn't closed under the consequence relation.

I cease to digress.

V. A theory of propositional attitudes should mesh with empirical accounts of mental processes.

We want a theory of PAs to say what (token) propositional attitudes *are;* or, at least, what the facts are in virtue of which PA ascriptions are true. It seems to me self-evident that no such theory could be acceptable unless it lent itself to explanations of the data—gross and commonsensical or subtle and experimental—about mental states and processes. This is not, of course, to require that a theory of PAs legitimize our current empirical psychology; only that it comport with some psychology or other that is independently warranted. I hear this as analogous to: the theory that water is H_2O couldn't be acceptable unless, taken together with appropriate empirical premises, it leads to explanations of the macro- and micro-properties of water. Hence, I hear it as undeniable.

I think, in fact, that the requirement that a theory of propositional attitudes should be empirically plausible can be made to do quite a lot of work; much more work than philosophers

have usually realized. I'll return to this presently, when we have some theories in hand.

Those, then, are the conditions that I want a theory of propositional attitudes to meet. I shall argue that, taken together, they strongly suggest that propositional attitudes are relations between organisms and formulae in an internal language; between organisms and internal sentences, as it were. It's convenient, however, to give the arguments in two steps; first, to show that conditions I-V comport nicely with the view that the objects of PAs are sentences, and then to show that these sentences are plausibly internal.

I begin by anticipating a charge of false advertising. The arguments to be reviewed are explicitly non-demonstrative. All I claim for the internal language theory is that it works (a) surprisingly well, and (b) better than any of the available alternatives. The clincher comes at the end: even if we didn't need internal sentences for purposes of I-V, we'd need them to do our psychology. Another non-demonstrative argument, no doubt, but one I find terrifically persuasive.

Carnap's Theory

Carnap suggested, in *Meaning and Necessity* (1947), that PAs might be construed as relations between people and sentences they are disposed to utter; e.g., between people and sentences of English. What Carnap had primarily in mind was coping with the opacity problem, but it's striking and instructive that his proposal does pretty well with *all* the conditions I've enumerated. Consider:

I. If propositional attitudes are relations to sentences, then they are relations *tout court*. Moreover, assume that the relation ascribed by a sentence of the form *a believes . . .* holds between the individual denoted by 'a' and the correspondent of the complement clause. It is then immediately clear why the belief ascribed to *a* is true if the correspondent is; the correspondent is the *object* of the belief (i.e., the correspondent is what's believed-true) if Carnap's story is right.

II. Vendler's condition is presumably satisfiable, though how the details go will depend on how we construe the objects of verbs of saying. A natural move for a neo-Carnapian to make would be to take 'John said that P' to be true in virtue of some relation between John and a token of the type P. Since, on this account, saying that P and believing that P involve relations to tokens of the very same sentence, it's hardly surprising that formulae which express the object of the *says-that* relation turn out to be logico-syntactically similar to formulae which express the object of the *believes-that* relation.

III. Frege's condition is satisfied; the opacity of belief is construed as a special case of the opacity of quotation. To put it slightly differently; 'John said 'Bill bit Mary" expresses a relation between John and a (quoted) sentence, so we're unsurprised by the fact that John may bear *that* relation to *that* sentence, while not bearing it to some arbitrarily similar but distinct sentence; e.g., to the sentence 'somebody bit Mary' or to the sentence 'Bill bit somebody,' etc. But ditto, *mutatis mutandis*, if 'John believes Bill bit Mary' *also* expresses a relation between John and a quoted sentence.

IV. Aristotle's condition is satisfied in the strong form. The logical form of the object of a belief sentence is inherited from the logical form of the correspondent of the belief sentence. Of course it is, since on the Carnap view, the correspondent of the belief sentence *is* the object of the belief that it ascribes.

V. Whether you think that Carnap's theory can claim empirical plausibility depends on what you take the empirical facts about propositional attitudes to be and how ingenious you are in exploiting the theory to provide explanations of the facts. Here's one example of how such an explanation might go.

It's plausible to claim that there is a fairly general parallelism between the complexity of beliefs and the complexity of the sentences that express them. So, for example, I take it that 'the Second Punic War was fought under conditions which neither of the combatants could have desired or foreseen' is a more complex sentence than, e.g., 'it's raining'; and, correspondingly, I take it that the thought that the Second Punic War was fought under conditions which neither of the combatants could have desired or foreseen is a more complicated thought than the thought that it's raining. Carnap's theory explains this parallelism[10] since, according to the theory, what makes a belief ascription true is a relation between an organism and the correspondent of

the belief-ascribing sentence. To hold the belief that the Second Punic War . . . , etc. is thus to be related to a more complex sentence than the one you are related to when you hold the belief that it's raining.

Some people need to count noses before they will admit to having one. In which case, see the discussion of 'codability' in Brown and Lenneberg 1954 and Brown 1976. What the experiments showed is that the relative complexity of the descriptions which subjects supply for color chips predicts the relative difficulty that the subjects have in identifying the chips in a recognition-recall task. Brown and Lenneberg explain the finding along strictly (though inadvertently) Carnapian lines: complex descriptions correspond to complex memories because it's the description which the subject (opaquely) remembers when he (transparently) remembers the color of the chip.

We can now begin to see *one* of the ways in which Condition V is supposed to work. A theory of propositional attitudes specifies a construal of the objects of the attitudes. It tells for such a theory if it can be shown to mesh with an independently plausible story about the 'cost accounting' for mental processes. A cost accounting function is just a (partial) ordering of mental states by their relative complexity. Such an ordering is, in turn, responsive to a variety of types of empirical data, both intuitive and experimental. Roughly, one has a 'mesh' between an empirically warranted cost accounting and a theory of the objects of PAs when one can predict the relative complexity of a mental state (or process) from the relative complexity of whatever the theory assigns as its object (or domain). (So, if Carnap is right, then the relative complexity of beliefs should be predictable from the relative linguistic complexity of the correspondents of belief ascribing sentences, all other things being equal.)

There's a good deal more to be said about all this than I have space for here. Again, roughly: to require that the complexity of the putative objects of PAs predict the cost accounting for the attitudes is to impose empirical constraints on the *notation* of (canonical) belief-ascribing sentences. So, for example, we would clearly get different predictions about the relative complexity of beliefs if we take the object of a PA to be the correspondent of the belief ascribing sentence than if we take it to be, e.g., the correspondent transformed into disjunctive form. The fact that there are empirical consequences of the notation we use to specify the objects of PAs is, of course, part and parcel of the fact that we are construing the attitude ascriptions *opaquely;* it is precisely under opaque construal that we distinguish (e.g.,) the mental state of believing that P & Q from the mental state of believing that neither not-P nor not-Q.

In short, Carnap's theory fares rather well with conditions I-V; there's more to be said in its favor than one might gather from the muted enthusiasm which philosophers have generally accorded it. Nevertheless, I think the philosophical consensus is warranted; Carnap's theory won't do. Here are some of the reasons.

1. Carnap has a theory about the objects of the propositional attitudes (viz., they're sentences) and a theory about the character of the relation to those objects in virtue of which one has a belief, desire, etc. Now, the latter theory is blatantly behavioristic; on Carnap's view, to believe that so-and-so is to be disposed (under presumably specifiable conditions) to utter tokens of the correspondent of the belief-ascribing sentence. But, patently, beliefs aren't behavioral dispositions; a fortiori, they aren't dispositions to utter. Hence, something's wrong with at least part of Carnap's account of the attitudes.

I put this objection first because it's the easiest to meet. So far as I can see, nothing prevents Carnap from keeping his account of the *objects* of belief while scuttling the behavioristic analysis of the belief relation. This would leave him wanting an answer to such questions as: what relation to the sentence 'it's raining' is such that you believe that it's raining if you and it are in that relation? In particular, he'd want some answer other than the behavioristic: 'It's the relation of being disposed to utter tokens of that sentence when. . . .'

The natural solution would be for Carnap to turn functionalist; to hold that to believe it's raining is to have a token of 'it's raining' play a certain role in the causation of your behavior and of your (other) mental states, said role eventually to be specified in the course of the detailed working out of empirical psychology . . . etc., etc. This is, perhaps, not much of a story, but it's fashionable, I know of nothing better, and it does have the virtue of explaining why propositional attitudes are opaque. Roughly, you wouldn't expect to be able to infer from 'tokens of the sentence S_1 have the causal role R' to 'tokens of the sentences S_2 have the causal role R' on the basis of any logical relation between

S_1 and S_2 (except, of course, identity). More generally, so far as I can see, a functionalist account of the way quoted sentences figure in the having of PAs will serve as well as a disposition-to-utter account in coping with all of conditions I-V. From now on, I'll take this emendation for granted.

2. The natural way to read the Carnap theory is to take type identity of the correspondents of belief ascribing sentences as necessary and sufficient for type identity of the ascribed beliefs; and it's at least arguable that this cuts the PAs too thin. So, for example, one might plausibly hold that 'John believes Mary bit Bill' and 'John believes Bill was bitten by Mary' ascribe the same belief (see note 9). In effect, this is the sinister side of the strategy of inheriting the opacity of belief from the opacity of quotation. The strategy fails whenever the identity conditions on beliefs are *different* from the identity conditions on sentences.

A way to cope would be to allow that the objects of beliefs are, in effect, *translation sets* of sentences; something like this seems to be the impetus for Carnap's doctrine of intentional isomorphism. In any event, the problems in this area are well-known. It may well be, for example, that the right way to characterize a translation relation for sentences is by referring to the communicative intentions of speaker/hearers of whatever language the sentences belong to. (S_1 translates S_2 if the two sentences are both standardly used with the same communicative intentions.) But, of course, we can't both identify translations by reference to intentions and individuate propositional attitudes (including, n.b., intentions) by reference to translations. This problem holds quite independent of epistemological worries about the facticity of ascriptions of propositional attitudes, the determinacy or otherwise of translations, etc.; which suggests that it may be serious.

3. You can believe that it's raining even if you don't speak English. This is a variant of the thickness of slice problem just mentioned; it again suggests that the appropriate objects of belief are translation sets and raises the specters that haunt that treatment.
4. You can, surely, believe that it's raining even if you don't speak any language at all. To say this is to say that at least *some* human cognitive psychology generalizes to infra-human organisms; if it didn't, we would find the behavior of animals *utterly* bewildering, which, in fact, we don't.

Of course, relations are cheap; there must be *some* relation which a dog bears to 'it's raining' if the dog believes that it's raining; albeit, perhaps, some not very interesting relation. So, why not choose *it* as the relation in virtue of which the belief-ascription holds of the dog? The problem is condition V. It would simply be a miracle if there were a relation between dogs and tokens of 'it's raining' such that any of the empirical facts about the propositional attitudinizing of dogs proved explicable in terms of that relation. (We can't, for example, choose any functional/causal relation because the behavior of dogs is surely not in any way caused by tokens of English sentences.) To put it generally if crudely, satisfying condition V depends on assuming that whatever the theory takes to be the object of a PA plays an appropriate role in the mental processes of the organism to which the attitude is ascribed. But English sentences play no role in the mental life of dogs. (Excepting, perhaps, such sentences as 'Down, Rover!' which, in any event, don't play the kind of role envisaged.)

5. We argued that the truth conditions on beliefs are inherited from the truth conditions on the correspondents of belief ascribing sentences, but this won't work if, for example, there are inexpressible beliefs. This problem is especially serious for behaviorist (or functionalist) accounts of the belief relation; to believe that P can't be a question of being disposed to utter (or of having one's behavior caused by) tokens of the sentence P if, as a matter of fact, there is no such sentence. Yet it is the appeal to quoted sentences which does the work in such theories: which allows them to satisfy I–V.
6. We remarked that there's a rough correspondence between the complexity of thoughts and the complexity of the sentences which express them, and that the (neo-) Carnapian theory provides for this; more generally, that the view that the objects of PAs are natural-language sentences might mesh reasonably well with an empirically defensible cost accounting for mental states and processes. Unfortunately this argument cuts both ways if we assume—as seems plausible—that the correspondence is no better than partial. Whenever it fails, there's prima

facie evidence *against* the theory that sentences are the objects of propositional attitudes.

In fact, we can do rather better than appealing to intuitions here. For example: we noted above that the 'codability' (viz., mean simplicity of descriptions in English) of colors predicts their recallability in a population of English-speakers, and that this comports with the view that what one remembers when one remembers a color is (at least sometimes) its description: i.e., with the view that descriptions are the objects of (at least some) propositional attitudes. It thus comes as a shock to find that codability *in English* also predicts recall for a Dani subject population. We can't explain this by assuming a correlation between codability-in-English and codability-in-Dani (i.e., by assuming that the colors that English speakers find easy to describe are the ones that Dani-speakers also find easy to describe) since, as it turns out, Dani has no vocabulary *at all* for chromatic variation; all such variation is *infinitely* uncodable in Dani. This comes close to being the paradox dreaded above: how could *English* sentences be the objects of the propositional attitudes of (monolingual) Dani? And, if they are not, how could a property defined over English sentences mesh with a theory of cost accounting for the mental processes of the Dani? It looks as though either: (a) some propositional attitudes are *not* relations to sentences, or (b) if they are—if English sentences are somehow the objects of Dani PAs—then sentences which constitute the objects of PAs need play no functional/causal role in the having of the attitudes. (For discussion of the cross-cultural results on codability, see Brown op. cit. For details of the original studies, see Heider 1972 and Berlin and Kay 1969.)

7. If (token) sentences of a natural language are the objects of propositional attitudes, how are (first) languages learned? On any theory of language learning we can now imagine that process must involve the collection of data, the formulation of hypotheses, the checking of the hypotheses against the data, and the decision about which of the hypotheses the data best confirm. That is, it must involve such mental states and processes as beliefs, expectation and perceptual integration. It's important to realize that *no* account of language learning which does not thus involve propositional attitudes and mental processes has ever been proposed by anyone, barring only behaviorists. And behaviorist accounts of language learning are, surely, not tenable. So, on pain of circularity, there must be *some* propositional attitudes which are not functional/causal relations to natural language sentences. I see no way out of this which isn't a worse option than rejecting the Carnap theory.

So, the situation looks discouraging. On the one hand, we have a number of plausible arguments in favor of accepting the Carnap story (viz., I–V) and, on the other, we have a number of equally plausible arguments in favor of not (viz. 1–7). Never mind; for, at second blush, it seems we needn't accept the whole Carnap theory to satisfy I–V and we needn't reject the whole Carnap theory to avoid 1–7. Roughly, all that I–V require is the part of the story that says that the objects of PAs are *sentences* (hence have logical forms, truth conditions, etc.). Whereas what causes the trouble with 1–7 is only that part of the story which says that they are *natural language* sentences (hence raising problems about non-verbal organisms, first language learning, etc.). The recommended solution is thus to take the objects of PAs to be sentences of a *non*-natural language; in effect, formulae in an Internal Representational System.

The first point is to establish that this proposal does what it is supposed to: copes with I–V without running afoul of 1–7. In fact, I propose to do less than that since, so far as I can see, the details would be extremely complicated. Suffice it here to indicate the general strategy.

Conditions I and III are relatively easy to meet. I demands that propositional attitudes be relations, and so they are if they are relations to internal representations. III demands a construal of opacity. Carnap met this demand by reducing the opacity of belief to the opacity of quotation, and so do we: the only difference is that, whereas for Carnap, 'John believes it's raining' relates John to a sentence of English, for us it relates John to an internal formula.

Conditions II and IV stress logico/syntactic parallelism between the complements and the correspondents of belief-ascribing sentences; such relations are epitomized by the identity between the truth conditions on 'it's raining' and those on what is believed when it's believed that it's raining. (Neo-) Carnap explained these symmetries by taking the correspondents of belief ascriptions to be the objects of beliefs. The present alternative is spiritually similar but one step less direct: we assume that the

correspondent of a belief-ascriber inherits its logico-semantic properties from the same internal formula which functions as the object of the belief ascribed.

There are three pieces in play: there are (a) *belief-ascribers* (like 'John believes it's raining'); (b) *complements* of belief ascribers (like ⌜it's raining⌝ in 'John believes it's raining'); and (c) *correspondents* of belief ascribers (like 'it's raining' standing free). The idea is to get all three to converge (though, of course, by different routes) on the same internal formula (call it 'F (it's raining)'[11]) thereby providing the groundwork for explaining the analogies that II and IV express.

To get this to work out right would be to supply detailed instructions for connecting the theory of PAs with the theory of sentence interpretation, and I have misplaced mine. But the general idea is apparent. Belief ascribers are true in virtue of functional/causal (call them 'belief making') relations between organisms and tokens of internal formulae. Thus, in particular, 'John believes it's raining' is true in virtue of a belief-making relation between John and a token of F (it's raining). It is, of course, the complement of a belief-ascriber that determines *which* internal formula is involved in its truth conditions; in effect 'it's raining' in 'John believes it's raining' functions as an index which picks out F (it's raining) and not, for example, F (elephants have wings) as the internal formula that John is related to if 'John believes it's raining' is true.

So, viewed along one vector, the complement of a belief-ascriber connects it with an internal formula. But, viewed along another vector, the complement of a belief ascriber connects it to its correspondent: if the correspondent of 'John believes it's raining' is 'it's raining,' that is because the form of words 'it's raining' constitutes its complement. And now we can close the circle, since, of course, F (it's raining) is *also* semantically connected with the correspondent of 'John believes it's raining' viz., by the principle that 'it's raining' is the sentence that English speakers use when they are in the belief-making relation to a token of F (it's raining) and wish to use a sentence of English to say what it is that they believe.

There are various ways of thinking about the relation between internal formulae and the correspondents of belief-ascribers. One is to think of the conventions of a natural language as functioning to establish a pairing of its verbal forms with the internal formulae that mediate the propositional attitudes of its users; in particular, as pairing the internal objects of beliefs with the form of words that speaker/hearers use to express their beliefs. This is a natural way to view the situation if you think of a natural language as a system of conventional vehicles for the expression of thoughts (a view to which I know of no serious objections). So in the present case, the conventions of English pair: 'it's raining' with F (it's raining) (viz., with the object of the belief that it's raining); 'elephants have wings' with F (elephants have wings) (viz., with the object of the belief that elephants have wings); and, generally, the object of each belief with the correspondent of some belief-ascribing sentence.[12]

Another option is to assume that F (it's raining) is distinguished by the fact that its tokens play a causal/functional role (not only as the object of the belief that it's raining, but also) in the production of linguistically regular utterances of 'it's raining.' Indeed, this option would plausibly be exercised in tandem with the one mentioned just above since it would be reasonable to construe 'linguistically regular' utterances as the ones that are produced in light of the speaker's knowledge of the linguistic conventions. The basic idea, in any event, would be to implicate F (it's raining) as the object of the communicative intentions that utterances of 'it's raining' standardly function to express; hence, as among the mental causes of such utterances. I take it that, given this relation, it ought to be possible to work out detailed tactics for the satisfaction of conditions II and IV, but this is the bit I propose to leave to the ingenuity of the reader. What I want to emphasize here is the way the linguistic structure of the complement of a belief ascriber connects it with free declaratives (in one direction) and with internal formulae (in the other). Contrary to the fusion story, it's no accident that 'it's raining' occurs in 'John believes it's raining.' Rather, the availability of natural languages for saying *both* what one believes *and* that one believes it turns on the exploitation of this elegant symmetry.

What about condition V? I shall consider this in conjunction with 2–7, since what's noteworthy about the latter is that they all register *empirical* complaints against the Carnap account. For example, 3, 4 and 6 would be without force if only everybody (viz., every subject of true propositional attitude ascriptions) talked English. 2 and 5 depend upon the empirical likelihood that English sentences fail

to correspond one-to-one to objects of propositional attitudes. 7 would be met if only English were innate. Indeed, I suppose an ultra-hard-line Neo-Carnapian might consider saving the bacon by claiming that—appearances to the contrary notwithstanding—English *is* innate, universal, just rich enough, etc. My point is that this is the right *kind* of move to make; all we have against it is its palpable untruth.

Whereas, it's part of the charm of the internal language story that, since practically nothing is known about the details of cognitive processes, we can make the corresponding assumptions about the internal representational system risking no more than gross implausibility at the very worst.

So, let's assume—what we don't, at any event, *know* to be false—that the internal language is innate, that its formulae correspond one-one with the contents of propositional attitudes (e.g., that 'John bit Mary' and 'Mary was bitten by John' correspond to the same 'internal sentence'), and that it is *as* universal as human psychology; viz., that to the extent that an organism shares our mental processes, it also shares our system of internal representations. On these assumptions, everything works. It's no longer paradoxical, for example, that codability *in English* predicts the relative complexity of the mental processes of the Dani; for, by assumption, it's not *really* the complexity of English sentences that predicts *our* cost accounting; we wouldn't expect *that* correspondence to be better than partial (see objection 6). What really predicts our cost accounting is the relative complexity of the internal representations that we use English sentences to express. And, again by assumption, the underlying system of internal representations is common to the Dani and to us. If you don't like this assumption, try and find some other hypothesis that accounts for the facts about the Dani.

Notice that to say that we can have our empirical assumptions isn't to say that we can have them for free. They carry a body of empirical commitments which, if untenable, will defeat the internal representation view. Imagine, for example, that cost accounting for English speakers proves utterly unrelated to cost accounting for (e.g.,) speakers of Latvian. (Imagine, in effect, that the Whorf-Sapir hypothesis turns out to be more or less true.) It's then hard to see how the system of internal representations could be universal. But if it's not universal, it's presumably not innate. And if it's not innate, it's not available to mediate the learning of first languages. And if it's not available to mediate the learning of first languages, we lose our means of coping with objection 7. There are plenty of ways in which we could find out that the theory's wrong if, in fact, it is.

Where we've gotten to is this: the general characteristics of propositional attitudes appear to demand sentence-like entities to be their objects. And broadly empirical conditions appear to preclude identifying these entities with sentences of *natural* languages; hence internal representations and private languages. How bad is it to have gotten here? I now want to argue that the present conclusion is independently required because it is presupposed by the best—indeed the only—psychology that we've got. Not just, as one philosopher has rather irresponsibly remarked, that 'some psychologists like to talk that way,' but that the best accounts of mental processes we have are quite unintelligible unless something like the internal representation story is true.

The long way of making this point is via a detailed discussion of such theories, but I've done that elsewhere and enough is enough. Suffice it here to consider a single example which is, however, prototypical. I claim again that the details don't matter; that one could make the same points by considering phenomena drawn from any area of cognitive psychology which is sufficiently well worked out to warrant talk of a theory *in situ.*

So, consider a fragment of contemporary (psycho)linguistics; consider the explanation of the ambiguity of a sentence like 'they are flying planes' (hereinafter, frequently *S*). The conventional story goes as follows: the sentence is ambiguous because there are two ways of grouping the word sequence into phrases, two ways of 'bracketing' it. One bracketing, corresponding to the reading of the sentence which answers 'what are those things?,' goes: (they) (are) (flying planes). Viz., the sentence is copular, the main verb is 'are' and 'flying' is an adjectival modifier of 'planes.' Whereas, on the other bracketing, corresponding to the reading on which the sentence answers 'what are those guys doing?,' the bracketing goes: (they) (are flying) (planes); viz. the sentence is transitive, the main verb is 'flying' and 'are' belongs to the auxiliary. I assume without argument that something like this is, or at least contributes to, the explanation of the ambiguity of *S*. The evidence for such treatments is overwhelming and there is, literally, no alternative theory in the field.

But what could it mean to speak of *S* as 'having' two bracketings? I continue to tread the well-worn path: *S* has two bracketings in that there exists a function (call it *G-proper*) from (as it might be) the word 'sentence' onto precisely those bracketed word strings which constitute the sentences of English. And both '(they) (are) (flying planes)' and '(they) (are flying) (planes)' are in the range of that function. (Moreover, no other bracketing of that word sequence is in the range of G-proper . . . etc.)

Now, the trouble with this explanation, as it stands, is that it is either enthymemic or silly. For, one wants to ask, how *could* the mere, as it were Platonic, existence of G-proper account for the facts about the ambiguity of English sentences? Or, to put it another way, sure there is, Platonically, a function under which *S* gets two bracketings. But there is also, Platonically, a function G′ under which it gets sixteen; and a function G″ under which it gets seven; and a function G‴ under which it gets none. Since G′, G″, and G‴ are all, qua functions, just as good as G-proper, how could the mere *existence* of the latter explain the linguistic properties of *S*? (You may feel inclined to say: 'Ah, but G-proper is the [or perhaps is *the*] grammar of English, and that distinguishes it from G′, G″ and the rest.' But this explanation takes one nowhere, since it invites the question: why does the grammar of English play a special role in the explanation of English sentences? Or, to put the same question minutely differently: call G′ the schmamar of English. We now want to know how come it's the bracketing assigned by English grammar and not the bracketing assigned by English schmamar, which predicts the ambiguity of 'they are flying planes'?)

So far as I can see, there's only one way such questions can conceivably be answered; viz., by holding that G-proper (not only exists but) is the very system of (internal [what else?]) formulae that English speaker/hearers use to represent the sentences of their language. But, then, if we accept this, we are willy-nilly involved in talking of at least *some* mental processes (processes of understanding and producing sentences) as involving at least some relations to at least some internal representations. And, if we have to have internal representations anyhow, why not take them to be the objects of propositional attitudes, thereby placating I–V? I say 'if we accept this'; but really we have no choice. For the account is well-evidenced, not demonstrably incoherent, and, again, it's the only one in the field. A working science is ipso facto in philosophical good repute.

So, by a series of non-demonstrative arguments: there are internal representations and propositional attitudes are relations that we bear to them. It remains to discuss two closely related objections.

Objection 1: Why not take the object of propositional attitudes to be *propositions?*

This suggestion has, no doubt, a ring of etymological plausibility; in fact, for all I know, it may be right. The mistake is in supposing it somehow conflicts with the present proposal.

I am taking seriously the idea that the system of internal representations constitutes a (computational) language. Qua language, it presumably has a syntax and a semantics; specifying the language involves saying what the properties are in virtue of which its formulae are well-formed, and what relations(s) obtain between the formulae and things in the (non-linguistic) world. I have no idea what an adequate semantics for a system of internal representations would look like; suffice it that, if propositions come in at all, they come in here. In particular, nothing stops us from specifying a semantics for the IRS by saying (inter alia) that some of its formulae express propositions. If we do say this, then we can make sense of the notion that propositional attitudes are relations to propositions; viz., they are *mediated* relations to propositions, with internal representations doing the mediating.

This is, quite generally, the way that representational theories of the mind work. So, in classical versions, thinking of John (construed opaquely) is a relation to an 'idea'; viz., to an internal representation of John. But this is quite compatible with its also being (transparently) construable as a relation *to John.* In particular, when Smith is thinking of John, he (normally) stands in relation to John and does so *in virtue* of his standing in relation to an idea of John. Similarly, mutatis mutandis, if thinking that it will rain is standing in relation to a proposition, then, on the present account, you stand in that relation in virtue of your (functional/causal) relation to an internal formula which expresses the proposition. No doubt, the 'expressing' bit is obscure; but that's a problem about propositions, not a problem about internal representations.

'Ah, but if you are going to allow propositions as the *mediate* objects of propositional attitudes, why bother with internal representations as their immediate objects? Why not just say: 'propositional attitudes are relations

to propositions. Punkt!" There's a small reason and a big reason. The small reason is that propositions don't have the right properties for our purposes. In particular, one anticipates problems of cost-accounting. Condition V, it will be remembered, permits us to choose among theories of PAs in virtue of the form of the entities they assign as objects of the attitudes. Now, the problem with propositions is that they are the sorts of things which, presumably, don't *have* forms. Propositions are sheer contents; they neutralize the lexico-syntactic differences between various ways of saying the same thing. That's what they're *for*. I say that this is a small problem but it looms prodigious if you hanker after a theory of the object of PAs which claims empirical repute. After all, it's not just cost-accounting which is supposed to be determined by formal aspects of the objects of PAs; it's *all* the mental processes and properties that cognitive psychology explains. That's what it *means* to speak of a *computational* psychology. Computational principles are ones that apply in virtue of the form of entities in their domain.

But my main reason for not saying 'propositional attitudes are relations to propositions. Punkt' is that I don't understand it. I don't see how an organism can stand in an (interesting epistemic) relation to a proposition except by standing in a (causal/functional) relation to some token of a formula which expresses the proposition. I am aware that there is a philosophical tradition to the contrary. Plato says (I think) that there is a special intellectual faculty (theoria) wherewith one peers at abstract objects. Frege says that one *apprehends* (what I'm calling) propositions, but I can find no doctrine about what apprehension comes to beyond the remark (in 'The Thought') that it's not sense perception because its objects are abstract and it's not introspection because its objects aren't mental. (He also says that grasping a thought isn't much like grasping a hammer. To be sure. As for me, I want a mechanism for the relation between organisms and propositions, and the only one I can think of is mediation by internal representations.[13])

Objection 2: Surely it's *conceivable* that propositional attitudes are *not* relations to internal representations.

I think it is; the theory that propositional attitudes are relations to internal representations is a piece of empirical psychology, not an analysis. For, there might have been angels, or behaviorism might have been true, and then the internal representation story would have been false. The moral is, I think, that we ought to give up asking for analyses; psychology is all the philosophy of mind that we are likely to get.

But, moreover, it may be *empirically* possible that there should be creatures which have the same propositional attitudes we do (e.g., the same beliefs) but *not* the same system of internal representations; creatures which, as it were, share our epistemic states but not our psychology. Suppose, for example, it turns out that Martians, or porpoises, believe what we do but have a very different sort of cost accounting. We might then want to say that there are translation relations among systems of internal representation (viz., that formally distinct representations can express the same proposition). Whether we can make sense of saying this remains to be seen; we can barely think about the question prior to the elaboration of theories about how such systems are semantically interpreted; and as things now stand, we haven't got semantic theories for natural languages, to say nothing of languages of thought. Perhaps it goes without saying that it's no objection to a doctrine that it *may* run us into incoherencies. Or, rather, if it is an objection, there's an adequate reply: 'Yes, but also it may not.'

I'll end on the note just sounded. Contemporary cognitive psychology is, in effect, a revival of the representational theory of the mind. The favored treatment of PAs arises in this context. So, in particular, the mind is conceived of as an organ whose function is the manipulation of representations and these, in turn, provide the domain of mental processes and the (immediate) objects of mental states. That's what it is to see the mind as something like a computer. (Or rather, to put the horse back in front of the cart, that's what it is to see a computer as something like the mind. We give sense to the analogy by treating selected states of the machine as formulae and by specifying which semantic interpretations the formulae are to bear. It is in the context of such specifications that we speak of machine processes as computations and of machine states as intensional.)

If the representational theory of the mind is true, then we know what propositional attitudes are. But the net total of philosophical problems is surely not decreased thereby. We must now face what has always been *the* problem for representational theories to solve: what relates internal representations to the world? What is it for a system of internal representations to be semantically interpreted? I take it that this problem is now the main content of the philosophy of mind.[14]

REFERENCES

Berlin, B. and Kay, P. *Basic Color Terms* (Berkeley: University of California Press, 1969).

Brown, R., "Reference—in memorial tribute to Eric Lenneberg," *Cognition* 4 (1976): pp. 125–53.

Brown, R., and Lenneberg, E., "A study in language and cognition," *Journal of Abnormal and Social Psychology* no. 49 (1954): pp. 454–62.

Carnap, R. *Meaning and Necessity* (Chicago: Phoenix Books, University of Chicago Press, 1947).

Church, A., "The need for abstract entities in semantic analysis, in *Contributions to the Analysis and Synthesis of Knowledge*, Proceedings of the American Academy of Arts and Sciences, no. 80 (1951): pp. 100–12.

Davidson, D., "Theories of meaning and learnable languages," in *Logic, Methodology and Philosophy of Science*, Proceedings of the 1964 International Congress, Y. Bar-Hillel, ed., (Amsterdam: North Holland Publishing Company, 1965), pp. 383–94.

Dennett, D. *Content and Consciousness* (New York: Routledge & Kegan Paul, 1969).

Dummett, M. *Frege* (London: Buckworth & Co., 1973).

Fodor, J. A. *The Language of Thought* (New York: Thomas Y. Crowell Co., 1975).

Goodman, N. *Languages of Art* (New York: Bobbs-Merrill, 1968).

Grice, H. P., "Method in philosophical psychology," in *Proceedings and Addresses of the American Philosophical Association*, vol. XLVIII (1975), pp. 23–53.

Heider, E., "Universals in color naming and memory," *Journal of Experimental Psychology*, no. 93 (1972): pp. 10–20.

Nagel, T., "Physicalism," *The Philosophical Review* 74 (1965): pp. 339–56.

Vendler, *Z. Res Cogitans* (Ithaca, NY: Cornell University Press, 1972).

NOTES

1. I shall have nothing at all to say about knowing, discovering, recognizing, or any other of the 'factive' attitudes. The justification for this restriction is worth discussing, but not here.
2. I haven't space to discuss here the idea that 'John believes' should be construed as an operator on 'it's raining.' Suffice it (a) that it's going to be hard to square that account with such observations as I-b below; and (b) that it seems quite implausible for such sentences as 'John believes what Mary said' (and what Mary said might *be* that it's raining). In general, the objects of propositional attitude verbs exhibit the syntax of object noun phrases, which is just what the operator account would not predict.
3. I assume that this is approximately correct: given a sentence of the syntactic form NP_1 (V [NP_2]) V expresses a relation if NP_1 and NP_2 refer. So, for present purposes, the question whether 'believes' expresses a relation in 'John believes it's raining' comes down to the question whether there are such things as objects of beliefs. I shan't, therefore, bother to distinguish among these various ways of putting the question in the discussion which follows.
4. Of course, it might refer in virtue of a relation between Mary and something other than a voice. 'John is taller than the average man' isn't true in virtue of a relation between John and the average man ('the average man' doesn't refer). But the sentence is relational for all that. It's for this sort of reason that such principles as the one announced in n3 hold only to a first approximation.
5. N.b., verbs of propositional attitude are transparent, in this sense, only when their objects are *complements;* one can't infer 'there is something Ponce de Leon sought' from 'Ponce de Leon sought the Fountain of Youth.' It may, however, be worth translating 'seek' to 'try to find' to save the generalization. This would give us: 'Ponce de Leon tried to find the Fountain of Youth,' which does, I suppose, entail that there is something that Ponce de Leon tried (viz., tried to do; viz., to find the Fountain of Youth).

   Also, to say that *EG* applies *to* the complement of verbs of PA is, of course, not to say that it applies *in* the complement of verbs of PA. 'John wants to marry Marie of Rumania' implies that there is something that John wants (viz., to marry Marie of Rumania); it notoriously does *not* imply that there is someone whom John wants to marry (see III below).
6. Fusion has been contemplated as a remedy for untransparency in several philosophical contexts; see Goodman 1968, Dennett 1969, Nagel 1965. N.b., 'contemplated,' not 'embraced.'
7. 3 is not a point about *EG.* On the fusion view, there's no representation of the fact that 'the belief that Sam is nice' is about Sam even when 'belief' and 'about' are both construed *opaquely.*
8. Defining 'correspondent' gets complicated where verbs of PA take *transformed* sentences as their objects, but the technicalities needn't concern us here. Suffice it that we want the correspondent of 'John wants to leave' to be 'John leaves'; the correspondent of 'John objects to Mary and Bill being elected' to be 'Mary and Bill are elected,' etc.
9. I am assuming that two sentences with correspondents of *different* logico-syntactic form cannot assign the same (opaque) belief, and someone might wish to challenge this; consider 'John believes that Mary bit Bill' and 'John believes that Bill was bitten by Mary.' This sort of objection is serious and will be accommodated later on.
10. In speaking of Carnap's theory, I don't wish to imply that Carnap would endorse the uses to which I'm putting it; quite the contrary, I should imagine.
11. Where *F* might be thought of as a function from (e.g., English) sentences onto internal formulae.
12. Assuming, as we may but now needn't do, that all beliefs are expressible in English. It is, of course, a consequence of the present view that all the beliefs we can entertain are expressible in the internal code.

13. The notion that the apprehension of propositions is mediated by linguistic objects is not entirely foreign even to the Platonistic tradition. Church says: '. . . the preference of (say) seeing over *understanding* as a method of observation seems to me capricious. For just as an opaque body may be seen, so a concept may be understood or grasped. . . . In both cases the observation is not direct but through intermediaries . . . linguistic expressions in the case of the concept' (1951a). See also the discussion in Dummett 1973, pp. 156–57.
14. All of the following helped: Professors Ned Block, Noam Chomsky, Dan Dennett, Hartry Field, Janet Dean Fodor, Keith Lehrer, and Brian Loar. Many thanks.

True Believers

The Intentional Strategy and Why It Works

Daniel C. Dennett

Death Speaks

> There was a merchant in Baghdad who sent his servant to market to buy provisions and in a little while the servant came back, white and trembling, and said, 'Master, just now when I was in the marketplace I was jostled by a woman in the crowd and when I turned I saw it was Death that jostled me. She looked at me and made a threatening gesture; now, lend me your horse, and I will ride away from this city and avoid my fate. I will go to Samarra and there Death will not find me.' The merchant lent him his horse, and the servant mounted it, and he dug his spurs in its flanks and as fast as the horse could gallop he went. Then the merchant went down to the marketplace and he saw me standing in the crowd, and he came to me and said, 'Why did you make a threatening gesture to my servant when you saw him this morning?' 'That was not a threatening gesture,' I said, 'it was only a start of surprise. I was astonished to see him in Baghdad, for I had an appointment with him tonight in Samarra.'
>
> W. Somerset Maugham

In the social sciences, talk about *belief* is ubiquitous. Since social scientists are typically self-conscious about their methods, there is also a lot of talk about *talk about belief*. And since belief is a genuinely curious and perplexing phenomenon, showing many different faces to the world, there is abundant controversy. Sometimes belief attribution appears to be a dark, risky, and imponderable business—especially when exotic, and more particularly religious or superstitious, beliefs are in the limelight. These are not the only troublesome cases; we also court argument and skepticism when we attribute beliefs to nonhuman animals, or to infants, or to computers or robots. Or when the beliefs we feel constrained to attribute to an apparently healthy, adult member of our own society are contradictory, or even just wildly false. A biologist colleague of mine was once called on the telephone by a man in a bar who wanted him to settle a bet. The man asked: 'Are rabbits birds?' 'No,' said the biologist. 'Damn!' said the man as he hung up. Now could he *really* have believed that rabbits were birds? Could anyone really and truly be attributed that belief? Perhaps, but it would take a bit of a story to bring us to accept it.

In all of these cases belief attribution appears beset with subjectivity, infected with cultural relativism, prone to 'indeterminacy of radical translation'—clearly an enterprise demanding special talents: the art of phenomenological analysis, hermeneutics, empathy, *Verstehen*, and all that. On other occasions, normal occasions, when familiar beliefs are the topic, belief attribution looks as easy as speaking prose and as objective and reliable as counting beans in a dish. Particularly when these straightforward cases are before us, it is quite plausible to suppose that in principle (if not yet in practice) it would be possible to confirm these simple, objective belief attributions by *finding something inside the believer's head*—by finding the beliefs themselves, in effect. 'Look,' someone

Reprinted from A. F. Heath, ed., *Scientific Explanations* (Oxford University Press, 1981), with permission of Oxford University Press.

might say, 'You either believe there's milk in the fridge or you don't believe there's milk in the fridge' (you might have no opinion, in the latter case). But if you do believe this, that's a perfectly objective fact about you, and it must come down in the end to your brain's being in some particular physical state. If we knew more about physiological psychology, we could in principle determine the facts about your brain state and thereby determine whether or not you believe there is milk in the fridge, even if you were determined to be silent or disingenuous on the topic. In principle, on this view physiological psychology could trump the results—or nonresults—of any 'black box' method in the social sciences that divines beliefs (and other mental features) by behavioral, cultural, social, historical, *external* criteria.

These differing reflections congeal into two opposing views on the nature of belief attribution, and hence on the nature of belief. The latter, a variety of *realism*, likens the question of whether a person has a particular belief to the question of whether a person is infected with a particular virus—a perfectly objective internal matter of fact about which an observer can often make educated guesses of great reliability. The former, which we could call *interpretationism* if we absolutely had to give it a name, likens the question of whether a person has a particular belief to the question of whether a person is immoral, or has style, or talent, or would make a good wife. Faced with such questions, we preface our answers with 'well, it all depends on what you're interested in,' or make some similar acknowledgment of the relativity of the issue. 'It's a matter of interpretation,' we say. These two opposing views, so baldly stated, do not fairly represent any serious theorists' positions, but they do express views that are typically seen as mutually exclusive and exhaustive; the theorist must be friendly with one and only one of these themes.

I think this is a mistake. My thesis will be that while belief is a perfectly objective phenomenon (that apparently makes me a realist), it can be discerned only from the point of view of one who adopts a certain *predictive strategy*, and its existence can be confirmed only by an assessment of the success of that strategy (that apparently makes me an interpretationist).

First I will describe the strategy, which I call the intentional strategy or adopting the intentional stance. To a first approximation, the intentional strategy consists of treating the object whose behavior you want to predict as a rational agent with beliefs and desires and other mental stages exhibiting what Brentano and others call *intentionality*. The strategy has often been described before, but I shall try to put this very familiar material in a new light by showing *how* it works and by showing *how well* it works.

Then I will argue that any object—or as I shall say, any *system*—whose behavior is well predicted by this strategy is in the fullest sense of the word a believer. *What it is* to be a true believer is to be an *intentional system*, a system whose behavior is reliably and voluminously predictable via the intentional strategy. I have argued for this position before (Dennett 1971, 1976, 1978), and my arguments have so far garnered few converts and many presumed counterexamples. I shall try again here, harder, and shall also deal with several compelling objections.

The Intentional Strategy and How It Works

There are many strategies, some good, some bad. Here is a strategy, for instance, for predicting the future behavior of a person: determine the date and hour of the person's birth and then feed this modest datum into one or another astrological algorithm for generating predictions of the person's prospects. This strategy is deplorably popular. Its popularity is deplorable only because we have such good reasons for believing that it does not work (*pace* Feyerabend 1978). When astrological predictions come true this is sheer luck, or the result of such vagueness or ambiguity in the prophecy that almost any eventuality can be construed to confirm it. But suppose the astrological strategy did in fact work well on some people. We could call those people *astrological systems*—systems whose behavior was, as a matter of fact, predictable by the astrological strategy. If there were such people, such astrological systems, we would be more interested than most of us in fact are in *how the astrological strategy works*—that is, we would be interested in the rules, principles, or methods of astrology. We could find out how the strategy works by asking astrologers, reading their books, and observing them in action. But we would also be curious about *why* it worked. We might find that astrologers had no useful opinions about this latter question—they either had no theory of why it worked or their theories were pure hokum. Having a good strategy is one thing; knowing why it works is another.

So far as we know, however, the class of astrological systems is empty, so the astrological strategy is of interest only as a social curiosity. Other strategies have better credentials. Consider the physical strategy, or physical stance; if you want to predict the behavior of a system, determine its physical constitution (perhaps all the way down to the microphysical level) and the physical nature of the impingements upon it, and use your knowledge of the laws of physics to predict the outcome for any input. This is the grand and impractical strategy of Laplace for predicting the entire future of everything in the universe, but it has more modest, local, actually usable versions. The chemist or physicist in the laboratory can use this strategy to predict the behavior of exotic materials, but equally the cook in the kitchen can predict the effect of leaving the pot on the burner too long. The strategy is not always practically available, but that it will always work *in principle* is a dogma of the physical sciences (I ignore the minor complications raised by the subatomic indeterminacies of quantum physics).

Sometimes, in any event, it is more effective to switch from the physical stance to what I call the design stance, where one ignores the actual (possibly messy) details of the physical constitution of an object, and, on the assumption that it has a certain design, predicts that it will behave *as it is designed to behave* under various circumstances. For instance, most users of computers have not the foggiest idea what physical principles are responsible for the computer's highly reliable, and hence predictable, behavior. But if they have a good idea of what the computer is designed to do (a description of its operation at any one of the many possible levels of abstraction), they can predict its behavior with great accuracy and reliability, subject to disconfirmation only in cases of physical malfunction. Less dramatically, almost anyone can predict when an alarm clock will sound on the basis of the most casual inspection of its exterior. One does not know or care to know whether it is spring wound, battery driven, sunlight powered, made of brass wheels and jewel bearings or silicon chips—one just assumes that it is designed so that the alarm will sound when it is set to sound, and it is set to sound where it appears to be set to sound, and the clock will keep on running until that time and beyond, and is designed to run more or less accurately, and so forth. For more accurate and detailed design stance predictions of the alarm clock, one must descend to a less abstract level of description of its design; for instance, to the level at which gears are described, but their material is not specified.

Only the designed behavior of a system is predictable from the design stance, of course. If you want to predict the behavior of an alarm clock when it is pumped full of liquid helium, revert to the physical stance. Not just artifacts but also many biological objects (plants and animals, kidneys and hearts, stamens and pistils) behave in ways that can be predicted from the design stance. They are not just physical systems but designed systems.

Sometimes even the design stance is practically inaccessible, and then there is yet another stance or strategy one can adopt: the intentional stance. Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many—but not all—instances yield a decision about what the agent ought to do; that is what you predict the agent *will* do.

The strategy becomes clearer with a little elaboration. Consider first how we go about populating each other's heads with beliefs. A few truisms: sheltered people tend to be ignorant; if you expose someone to something he comes to know all about it. In general, it seems, we come to believe all the truths about the parts of the world around us we are put in a position to learn about. Exposure to x, that is, sensory confrontation with x over some suitable period of time, is the *normally sufficient* condition for knowing (or having true beliefs) about x. As we say, we come to *know all about* the things around us. Such exposure is only *normally* sufficient for knowledge, but this is not the large escape hatch it might appear; our threshold for accepting abnormal ignorance in the face of exposure is quite high. 'I didn't know the gun was loaded,' said by one who was observed to be present, sighted, and awake during the loading, meets with a variety of utter skepticism that only the most outlandish supporting tale could overwhelm.

Of course we do not come to learn or remember all the truths our sensory histories avail us. In spite of the phrase 'know all about,' what we come to know, normally, are only all the

relevant truths our sensory histories avail us. I do not typically come to know the ratio of spectacle-wearing people to trousered people in a room I inhabit, though if this interested me, it would be readily learnable. It is not just that some facts about my environment are below my thresholds of discrimination or beyond the integration and holding power of my memory (such as the height in inches of all the people present), but that many perfectly detectable, graspable, memorable facts are of no interest to me and hence do not come to be believed by me. So one rule for attributing beliefs in the intentional strategy is this: attribute as beliefs all the truths relevant to the system's interests (or desires) that the system's experience to date has made available. This rule leads to attributing somewhat too much—since we all are somewhat forgetful, even of important things. It also fails to capture the false beliefs we are all known to have. But the attribution of false belief, *any* false belief, requires a special genealogy, which will be seen to consist in the main in true beliefs. Two paradigm cases: *S* believes (falsely) that *p*, because *S* believes (truly) that Jones told him that *p*, that Jones is pretty clever, that Jones did not intend to deceive him, . . . etc. Second case: *S* believes (falsely) that there is a snake on the barstool, because *S* believes (truly) that he seems to see a snake on the barstool, is himself sitting in a bar not a yard from the barstool he sees, and so forth. The falsehood has to start somewhere; the seed may be sown in hallucination, illusion, a normal variety of simple misperception, memory deterioration, or deliberate fraud, for instance, but the false beliefs that are reaped grow in a culture medium of true beliefs.

Then there are the arcane and sophisticated beliefs, true and false, that are so often at the focus of attention in discussions of belief attribution. They do not arise directly, goodness knows, from exposure to mundane things and events, but their attribution requires tracing out a lineage of mainly good argument or reasoning from the bulk of beliefs already attributed. An implication of the intentional strategy, then, is that true believers mainly believe truths. If anyone could devise an agreed-upon method of individuating and counting beliefs (which I doubt very much), we would see that all but the smallest portion (say, less than ten percent) of a person's beliefs were attributable under our first rule.[1]

Note that this rule is a derived rule, an elaboration and further specification of the fundamental rule: attribute those beliefs the system *ought to have*. Note also that the rule interacts with the attribution of desires. How do we attribute the desires (preferences, goals, interests) on whose basis we will shape the list of beliefs? We attribute the desires the system *ought to have*. That is the fundamental rule. It dictates, on a first pass, that we attribute the familiar list of highest, or most basic, desires to people: survival, absence of pain, food, comfort, procreation, entertainment. Citing any one of these desires typically terminates the 'Why?' game of reason giving. One is not supposed to need an ulterior motive for desiring comfort or pleasure or the prolongation of one's existence. Derived rules of desire attribution interact with belief attributions. Trivially, we have the rule: attribute desires for those things a system believes to be good for it. Somewhat more informatively, attribute desires for those things a system believes to be best means to other ends it desires. The attribution of bizarre and detrimental desires thus requires, like the attribution of false beliefs, special stories.

The interaction between belief and desire becomes trickier when we consider what desires we attribute on the basis of verbal behavior. The capacity to *express* desires in language opens the floodgates of desire attribution. 'I want a two-egg mushroom omelette, some French bread and butter, and a half bottle of lightly chilled white Burgundy.' How could one begin to attribute a desire for anything so specific in the absence of such verbal declaration? How, indeed, could a creature come to *contract* such a specific desire without the aid of language? Language *enables* us to formulate highly specific desires, but it also *forces* us on occasion to commit ourselves to desires altogether more stringent in their conditions of satisfaction than anything we would otherwise have any reason to endeavor to satisfy. Since in order to get what you want you often have to say what you want, and since you often cannot say what you want without saying something more specific than you antecedently mean, you often end up giving others evidence—the very best of evidence, your unextorted word—that you desire things or states of affairs far more particular than would satisfy you—or better, than would have satisfied you, for once you have declared, being a man of your word, you acquire an interest in satisfying exactly the desire you declared and no other.

'I'd like some baked beans, please.'

'Yes, sir. How many?'

You might well object to having such a specification of desire demanded of you, but in fact we are all socialized to accede to similar requirements in daily life—to the point of not noticing it, and certainly not feeling oppressed by it. I dwell on this because it has a parallel in the realm of belief, where our linguistic environment is forever forcing us to give—or concede—precise verbal expression to convictions that lack the hard edges verbalization endows them with (see Dennett 1969, pp. 184–85, and Dennett 1978, chapter 16). By concentrating on the *results* of this social force, while ignoring its distorting effect, one can easily be misled into thinking that it is *obvious* that beliefs and desires are rather like sentences stored in the head. Being language-using creatures, it is inevitable that we should often come to believe that some particular, actually formulated, spelled and punctuated sentence *is true*, and that on other occasions we should come to want such a sentence to *come true*, but these are special cases of belief and desire and as such may not be reliable models for the whole domain.

That is enough, on this occasion, about the principles of belief and desire attribution to be found in the intentional strategy. What about the rationality one attributes to an intentional system? One starts with the ideal of perfect rationality and revises downward as circumstances dictate. That is, one starts with the assumption that people believe all the implications of their beliefs and believe no contradictory pairs of beliefs. This does not create a practical problem of clutter (infinitely many implications, for instance), for one is interested only in ensuring that the system one is predicting is rational enough to get to the particular implications that are relevant to its behavioral predicament of the moment. Instances of irrationality, or of finitely powerful capacities of inferences, raise particularly knotty problems of interpretation, which I will set aside on this occasion (see Dennett 1987, chapter 4, and Cherniak 1986).

For I want to turn from the description of the strategy to the question of its use. Do people actually use this strategy? Yes, all the time. There may someday be other strategies for attributing belief and desire and for predicting behavior, but this is the only one we all know now. And when does it work? It works with people almost all the time. Why would it *not* be a good idea to allow individual Oxford colleges to create and grant academic degrees whenever they saw fit? The answer is a long story, but very easy to generate. And there would be widespread agreement about the major points. We have no difficulty thinking of the reasons people would then have for acting in such ways as to give others reasons for acting in such ways as to give others reasons for . . . creating a circumstance we would not want. Our use of the intentional strategy is so habitual and effortless that the role it plays in shaping our expectations about people is easily overlooked. The strategy also works on most other mammals most of the time. For instance, you can use it to design better traps to catch those mammals, by reasoning about what the creature knows or believes about various things, what it prefers, what it wants to avoid. The strategy works on birds, and on fish, and on reptiles, and on insects and spiders, and even on such lowly and unenterprising creatures as clams (once a clam believes there is danger about, it will not relax its grip on its closed shell until it is convinced that the danger has passed). It also works on some artifacts: the chess-playing computer will not take your knight because it knows that there is a line of ensuing play that would lead to losing its rook, and it does not want that to happen. More modestly, the thermostat will turn off the boiler as soon as it comes to believe the room has reached the desired temperature.

The strategy even works for plants. In a locale with late spring storms, you should plant apple varieties that are particularly *cautious* about *concluding* that it is spring—which is when they *want* to blossom, of course. It even works for such inanimate and apparently undesigned phenomena as lightning. An electrician once explained to me how he worked out how to protect my underground water pump from lightning damage: lightning, he said, always wants to find the best way to ground, but sometimes it gets tricked into taking second-best paths. You can protect the pump by making another, better path more *obvious* to the lightning.

True Believers as Intentional Systems

Now clearly this is a motley assortment of 'serious' belief attributions, dubious belief attributions, pedagogically useful metaphors, *façons de parler*, and, perhaps worse, outright frauds. The next task would seem to be distinguishing those intentional systems that *really* have beliefs and desires from those we may find it handy to treat *as if* they had beliefs and

desires. But that would be a Sisyphean labor, or else would be terminated by fiat. A better understanding of the phenomenon of belief begins with the observation that even in the worst of these cases, even when we are surest that the strategy works *for the wrong reasons*, it is nevertheless true that it does work, at least a little bit. This is an interesting fact, which distinguishes this class of objects, the class of *intentional systems*, from the class of objects for which the strategy never works. But is this so? Does our definition of an intentional system exclude any objects at all? For instance, it seems the lectern in this lecture room can be construed as an intentional system, fully rational, believing that it is currently located at the center of the civilized world (as some of you may also think), and desiring above all else to remain at that center. What should such a rational agent so equipped with belief and desire do? Stay put, clearly, which is just what the lectern does. I predict the lectern's behavior, accurately, from the intentional stance, so is it an intentional system? If it is, anything at all is.

What should disqualify the lectern? For one thing, the strategy does not recommend itself in this case, for we get no predictive power from it that we did not antecedently have. We already knew what the lectern was going to do—namely nothing—and tailored the beliefs and desires to fit in a quite unprincipled way. In the case of people or animals or computers, however, the situation is different. In these cases often the only strategy that is at all practical is the intentional strategy; it gives us predictive power we can get by no other method. But, it will be urged, this is no difference in nature, but merely a difference that reflects upon our limited capacities as scientists. The Laplacean omniscient physicist could predict the behavior of a computer—or of a live human body, assuming it to be ultimately governed by the laws of physics—without any need for the risky, short-cut methods of either the design or intentional strategies. For people of limited mechanical aptitude, the intentional interpretation of a simple thermostat is a handy and largely innocuous crutch, but the engineers among us can quite fully grasp its internal operation without the aid of this anthropomorphizing. It may be true that the cleverest engineers find it practically impossible to maintain a clear conception of more complex systems, such as a time-sharing computer system or remote-controlled space probe, without lapsing into an intentional stance (and viewing these devices as asking and telling, trying and avoiding, wanting and believing), but this is just a more advanced case of human epistemic frailty. We would not want to classify these artifacts with the true believers—ourselves—on such variable and parochial grounds, would we? Would it not be intolerable to hold that some artifact or creature or person was a believer from the point of view of one observer, but not a believer at all from the point of view of another, cleverer observer? That would be a particularly radical version of interpretationism, and some have thought I espoused it in urging that belief be viewed in terms of the success of the intentional strategy. I must confess that my presentation of the view has sometimes invited that reading, but I now want to discourage it. The decision to adopt the intentional stance is free, but the facts about the success or failure of the stance, were one to adopt it, are perfectly objective.

Once the intentional strategy is in place, it is an extraordinarily powerful tool in prediction—a fact that is largely concealed by our typical concentration on the cases in which it yields dubious or unreliable results. Consider, for instance, predicting moves in a chess game. What makes chess an interesting game, one can see, is the *un*predictability of one's opponent's moves, except in those cases where moves are 'forced'—where there is *clearly* one best move—typically the least of the available evils. But this unpredictability is put in context when one recognizes that in the typical chess situation there are very many perfectly legal and hence available moves, but only a few—perhaps half a dozen—with anything to be said for them, and hence only a few high-probability moves according to the intentional strategy. Even when the intentional strategy fails to distinguish a single move with a highest probability, it can dramatically reduce the number of live options.

The same feature is apparent when the intentional strategy is applied to 'real world' cases. It is notoriously unable to predict the exact purchase and sell decisions of stock traders, for instance, or the exact sequence of words a politician will utter when making a scheduled speech, but one's confidence can be very high indeed about slightly less specific predictions: that the particular trader *will not buy utilities today*, or that the politician *will side with the unions against his party*, for example. This inability to predict fine-grained descriptions of actions, looked at another way, is a source of strength for the intentional strategy, for it is this neutrality with regard to details of implementation that permits one to exploit the

intentional strategy in complex cases, for instance, in *chaining predictions* (see Dennett 1978). Suppose the US Secretary of State were to announce he was a paid agent of the KGB. What an unparalleled event! How unpredictable its consequences! Yet in fact we can predict dozens of not terribly interesting but perfectly salient consequences, and consequences of consequences. The President would confer with the rest of the Cabinet, which would support his decision to relieve the Secretary of State of his duties pending the results of various investigations, psychiatric and political, and all this would be reported at a news conference to people who would write stories that would be commented upon in editorials that would be read by people who would write letters to the editors, and so forth. None of that is daring prognostication, but note that it describes an arc of causation in space-time that could not be predicted under *any* description by any imaginable practical extension of physics or biology.

The power of the intentional strategy can be seen even more sharply with the aid of an objection first raised by Robert Nozick some years ago. Suppose, he suggested, some beings of vastly superior intelligence—from Mars, let us say—were to descend upon us, and suppose that we were to them as simple thermostats are to clever engineers. Suppose, that is, that they did not *need* the intentional stance—or even the design stance—to predict our behavior in all its detail. They can be supposed to be Laplacean super-physicists, capable of comprehending the activity on Wall Street, for instance, at the microphysical level. Where we see brokers and buildings and sell orders and bids, they see vast congeries of subatomic particles milling about—and they are such good physicists that they can predict days in advance what ink marks will appear each day on the paper tape labelled 'Closing Dow Jones Industrial Average.' They can predict the individual behaviors of all the various moving bodies they observe without ever treating any of them as intentional systems. Would we be right then to say that from *their* point of view we really were not believers at all (any more than a simple thermostat is)? If so, then our status as believers is nothing objective, but rather something in the eye of the beholder—provided the beholder shares our intellectual limitations.

Our imagined Martians might be able to predict the future of the human race by Laplacean methods, but if they did not also see us as intentional systems, they would be missing something perfectly objective: the *patterns* in human behavior that are describable from the intentional stance, and only from that stance, and that support generalizations and predictions. Take a particular instance in which the Martians observe a stockbroker deciding to place an order for 500 shares of General Motors. They predict the exact motions of his fingers as he dials the phone and the exact vibrations of his vocal cords as he intones his order. But if the Martians do not see that indefinitely many *different* patterns of finger motions and vocal cord vibrations—even the motions of indefinitely many different individuals—could have been substituted for the actual particulars without perturbing the subsequent operation of the market, then they have failed to see a real pattern in the world they are observing. Just as there are indefinitely many ways of *being a spark plug*—and one has not understood what an internal combustion engine is unless one realizes that a variety of different devices can be screwed into these sockets without affecting the performance of the engine—so there are indefinitely many ways of *ordering 500 shares of General Motors*, and there are societal sockets in which one of these ways will produce just about the same effect as any other. There are also societal pivot points, as it were, where which way people go depends on whether they *believe that p*, or *desire A*, and does not depend on any of the other infinitely many ways they may be alike or different.

Suppose, pursuing our Martian fantasy a little further, that one of the Martians were to engage in a predicting contest with an Earthling. The Earthling and the Martian observe (and observe each other observing) a particular bit of local physical transaction. From the Earthling's point of view, this is what is observed. The telephone rings in Mrs. Gardner's kitchen. She answers, and this is what she says: 'Oh, hello dear. You're coming home early? Within the hour? And bringing the boss to dinner? Pick up a bottle of wine on the way home, then, and drive carefully.' On the basis of this observation, our Earthling predicts that a large metallic vehicle with rubber tires will come to a stop in the drive within one hour, disgorging two human beings, one of whom will be holding a paper bag containing a bottle containing an alcoholic fluid. The prediction is a bit risky, perhaps, but a good bet on all counts. The Martian makes the same prediction, but has to avail himself of much more information about an extraordinary number of interactions of which, so far as he can tell, the Earthling

is entirely ignorant. For instance, the deceleration of the vehicle at intersection *A*, five miles from the house, without which there would have been a collision with another vehicle—whose collision course had been laboriously calculated over some hundreds of meters by the Martian. The Earthling's performance would look like magic! How did the Earthling know that the human being who got out of the car and got the bottle in the shop would get back in? The coming true of the Earthling's prediction, after all the vagaries, intersections, and branches in the paths charted by the Martian, would seem to anyone bereft of the intentional strategy as marvelous and inexplicable as the fatalistic inevitability of the appointment in Samarra. Fatalists—for instance, astrologers—believe that there is a pattern in human affairs that is inexorable, that will impose itself *come what may*, that is, no matter how the victims scheme and second-guess, no matter how they twist and turn in their chains. These fatalists are wrong, but they are *almost* right. There *are* patterns in human affairs that impose themselves, not quite inexorably but with great vigor, absorbing physical perturbations and variations that might as well be considered random; these are the patterns that we characterize in terms of the beliefs, desires, and intentions of rational agents.

No doubt you will have noticed, and been distracted by, a serious flaw in our thought experiment: the Martian is presumed to treat his Earthling opponent as an intelligent being like himself, with whom communication is possible, a being with whom one can make a wager, against whom one can compete. In short, a being with beliefs (such as the belief he expressed in his prediction) and desires (such as the desire to win the prediction contest). So if the Martian sees the pattern in one Earthling, how can he fail to see it in the others? As a bit of narrative, our example could be strengthened by supposing that our Earthling cleverly learned Martian (which is transmitted by X-ray modulation) and disguised himself as a Martian, counting on the species-chauvinism of these otherwise brilliant aliens to permit him to pass as an intentional system while not giving away the secret of his fellow human beings. This addition might get us over a bad twist in the tale, but might obscure the moral to be drawn: namely, *the unavoidability of the intentional stance with regard to oneself and one's fellow intelligent beings.* This unavoidability is itself interest relative; it is perfectly possible to adopt a physical stance, for instance, with regard to an intelligent being, oneself included, but not to the exclusion of maintaining at the same time an intentional stance with regard to oneself at a minimum, and one's fellows *if* one intends, for instance, to learn what they know (a point that has been powerfully made by Stuart Hampshire in a number of writings). We can perhaps suppose our super-intelligent Martians fail to recognize *us* as intentional systems, but we cannot suppose them to lack the requisite concepts.[2] If they observe, theorize, predict, communicate, they view *themselves* as intentional systems.[3] Where there are intelligent beings, the patterns must be there to be described, whether or not we care to see them.

It is important to recognize the objective reality of the intentional patterns discernible in the activities of intelligent creatures, but also important to recognize the incompleteness and imperfections in the patterns. The objective fact is that the intentional strategy *works as well as it does*, which is not perfectly. No one is perfectly rational, perfectly unforgetful, all-observant, or invulnerable to fatigue, malfunction, or design imperfection. This leads inevitably to circumstances beyond the power of the intentional strategy to describe, in much the same way that physical damage to an artifact, such as a telephone or an automobile, may render it indescribable by the normal design terminology for that artifact. How do you draw the schematic wiring diagram of an audio amplifier that has been partially melted, or how do you characterize the program state of a malfunctioning computer? In cases of even the mildest and most familiar cognitive pathology—where people seem to hold contradictory beliefs or to be deceiving themselves, for instance—the canons of interpretation of the intentional strategy fail to yield clear, stable verdicts about which beliefs and desires to attribute to a person.

Now a *strong* realist position on beliefs and desires would claim that in these cases the person in question really does have some particular beliefs and desires which the intentional strategy, as I have described it, is simply unable to divine. On the milder sort of realism I am advocating, there is no fact of the matter of exactly which beliefs and desires a person has in these degenerate cases, but this is not a surrender to relativism or subjectivism, for *when* and *why* there is no fact of the matter is itself a matter of objective fact. On this view one can even acknowledge the *interest relativity* of belief attributions and grant that given the different interests of different cultures, for instance, the beliefs and desires one culture would attribute to a member

might be quite different from the beliefs and desires another culture would attribute to that very same person. But supposing that were so in a particular case, there would be the further facts about *how well* each of the rival intentional strategies worked for predicting the behavior of that person. We can be sure in advance that no intentional interpretation of an individual will work to perfection, and it may be that two rival schemes are about equally good, and better than any others we can devise. That this is the case is itself something about which there can be a fact of the matter. The objective presence of one pattern (with whatever imperfections) does not rule out the objective presence of another pattern (with whatever imperfections).

The bogey of radically different interpretations with equal warrant from the intentional strategy is theoretically important—one might better say metaphysically important—but practically negligible once one restricts one's attention to the largest and most complex intentional systems we know: human beings.[4]

Until now I have been stressing our kinship to clams and thermostats, in order to emphasize a view of the logical status of belief attribution, but the time has come to acknowledge the obvious differences and say what can be made of them. The perverse claim remains: *all there is* to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy, and hence *all there is* to really and truly believing that *p* (for any proposition *p*) is being an intentional system for which *p* occurs as a belief in the best (most predictive) interpretation. But once we turn our attention to the truly interesting and versatile intentional systems, we see that this apparently shallow and instrumentalistic criterion of belief puts a severe constraint on the internal constitution of a genuine believer, and thus yields a robust version of belief after all.

Consider the lowly thermostat, as degenerate a case of an intentional system as could conceivably hold our attention for more than a moment. Going along with the gag, we might agree to grant it the capacity for about half a dozen different beliefs and fewer desires—it can believe the room is too cold or too hot, that the boiler is on or off, and that if it wants the room warmer it should turn on the boiler, and so forth. But surely this is imputing too much to the thermostat; it has no concept of heat or of a boiler, for instance. So suppose we *de-interpret* its beliefs and desires: it can believe the *A* is too *F* or *G*, and if it wants the *A* to be more *F* it should do *K*, and so forth. After all, by attaching the thermostatic control mechanism to different input and output devices, it could be made to regulate the amount of water in a tank, or the speed of a train, for instance. Its attachment to a heat-sensitive transducer and a boiler is too impoverished a link to the world to grant any rich semantics to its belief-like states.

But suppose we then enrich these modes of attachment. Suppose we give it more than one way of learning about the temperature, for instance. We give it an eye of sorts that can distinguish huddled, shivering occupants of the room and an ear so that it can be told how cold it is. We give it some facts about geography so that it can conclude that it is probably in a cold place if it learns that its spatiotemporal location is Winnipeg in December. Of course giving it a visual system that is multipurpose and general—not a mere shivering-object detector—will require vast complications of its inner structure. Suppose we also give our system more behavioral versatility: it chooses the boiler fuel, purchases it from the cheapest and most reliable dealer, checks the weather stripping, and so forth. This adds another dimension of internal complexity; it gives individual belief-like states *more to do*, in effect, by providing more and different occasions for their derivation or deduction from other states, and by providing more and different occasions for them to serve as premises for further reasoning. The cumulative effect of enriching these connections between the device and the world in which it resides is to enrich the semantics of its dummy predicates, *F* and *G* and the rest. The more of this we add, the less amenable our device becomes to serving as the control structure of anything other than a room-temperature maintenance system. A more formal way of saying this is that the class of indistinguishably satisfactory models of the formal system embodied in its internal states gets smaller and smaller as we add such complexities; the more we add, the richer or more demanding or specific the semantics of the system, until eventually we reach systems for which a unique semantic interpretation is practically (but never in principle) dictated (cf. Hayes 1979). At that point we say this device (or animal or person) has beliefs *about heat* and *about this very room*, and so forth, not only because of the system's actual location in, and operations on, the world, but because we cannot imagine another niche in which it could be placed *where it would work* (see also Dennett 1987, chapters 5 and 8).

Our original simple thermostat had a state we called a belief about a particular boiler, to the effect that it was on or off. Why about *that* boiler? Well, what other boiler would you want to say it was about? The belief is about the boiler because it is *fastened* to the boiler.[5] Given the actual, if mimimal, causal link to the world that happened to be in effect, we could endow a state of the device with *meaning* (of a sort) and *truth conditions*, but it was altogether too easy to substitute a different minimal link and completely change the meaning (in this impoverished sense) of that internal state. But as systems become perceptually richer and behaviorally more versatile, it becomes harder and harder to make substitutions in the actual links of the system to the world without changing the organization of the system itself. If you change its environment, it will *notice*, in effect, and make a change in its internal state in response. There comes to be a two-way constraint of growing specificity between the device and the environment. Fix the device in any one state and it demands a very specific environment in which to operate properly (you can no longer switch it easily from regulating temperature to regulating speed or anything else); but at the same time, if you do not *fix* the state it is in, but just plonk it down in a changed environment, its sensory attachments will be sensitive and discriminative enough to respond appropriately to the change, driving the system into a new state, in which it will operate effectively in the new environment. There is a familiar way of alluding to this tight relationship that can exist between the organization of a system and its environment: you say that the organism continuously *mirrors* the environment, or that there is a *representation* of the environment in—or implicit in—the organization of the system.

It is not that we attribute (or should attribute) beliefs and desires only to things in which we find internal representations, but rather that when we discover some object for which the intentional strategy works, we endeavor to interpret some of its internal states or processes as internal representations. What makes some internal feature of a thing a representation could only be its role in regulating the behavior of an intentional system.

Now the reason for stressing our kinship with the thermostat should be clear. There is no magic moment in the transition from a simple thermostat to a system that *really* has an internal representation of the world around it. The thermostat has a minimally demanding representation of the world, fancier thermostats have more demanding representations of the world, fancier robots for helping around the house would have still more demanding representations of the world. Finally you reach us. We are so multifariously and intricately connected to the world that almost no substitution is possible—though it is clearly imaginable in a thought experiment. Hilary Putnam imagines the planet Twin Earth, which is just like Earth right down to the scuff marks on the shoes of the Twin Earth replica of your neighbor, but which differs from Earth in some property that is entirely beneath the thresholds of your capacities to discriminate. (What they call water on Twin Earth has a different chemical analysis.) Were *you* to be whisked instantaneously to Twin Earth and exchanged for your Twin Earth replica, you would never be the wiser—just like the simple control system that cannot tell whether it is regulating temperature, speed, or volume of water in a tank. It is easy to devise radically different Twin Earths for something as simple and sensorily deprived as a thermostat, but your internal organization puts a much more stringent demand on substitution. Your Twin Earth and Earth must be virtual replicas or you will change state dramatically on arrival.

So which boiler are *your* beliefs about when you believe the boiler is on? Why, the boiler in your cellar (rather than its twin on Twin Earth, for instance). What other boiler would your beliefs be about? The completion of the semantic interpretation of your beliefs, fixing the referents of your beliefs, requires, as in the case of the thermostat, facts about your actual embedding in the world. The principles, and problems, of interpretation that we discover when we attribute beliefs to people are the *same* principles and problems we discover when we look at the ludicrous, but blessedly simple, problem of attributing beliefs to a thermostat. The differences are of degree, but nevertheless of such great degree that understanding the internal organization of a simple intentional system gives one very little basis for understanding the internal organization of a complex intentional system, such as a human being.

Why Does the Intentional Strategy Work?

When we turn to the question of *why* the intentional strategy works as well as it does, we find that the question is ambiguous, admitting of

two very different sorts of answers. If the intentional system is a simple thermostat, one answer is simply this: the intentional strategy works because the thermostat is well designed; it was designed to be a system that could be easily and reliably comprehended and manipulated from this stance. That is true, but not very informative, if what we are after are the actual features of its design that explain its performance. Fortunately, however, in the case of a simple thermostat those features are easily discovered and understood, so the other answer to our *why* question, which is really an answer about *how the machinery works*, is readily available.

If the intentional system in question is a person, there is also an ambiguity in our question. The first answer to the question of why the intentional strategy works is that evolution has designed human beings to be rational, to believe what they ought to believe and want what they ought to want. The fact that we are products of a long and demanding evolutionary process guarantees that using the intentional strategy on us is a safe bet. This answer has the virtues of truth and brevity, and on this occasion the additional virtue of being an answer Herbert Spencer would applaud, but it is also strikingly uninformative. The more difficult version of the question asks, in effect, how the machinery which Nature has provided us works. And we cannot yet give a good answer to that question. We just do not know. We do know how the *strategy* works, and we know the easy answer to the question of why it works, but knowing these does not help us much with the hard answer.

It is not that there is any dearth of doctrine, however. A Skinnerian behaviorist, for instance, would say that the strategy works because its imputations of beliefs and desires are short-hand, in effect, for as yet unimaginably complex descriptions of the effects of prior histories of response and reinforcement. To say that someone wants some ice cream is to say that in the past the ingestion of ice cream has been reinforced in him by the results, creating a propensity under certain background conditions (also too complex to describe) to engage in ice-cream-acquiring behavior. In the absence of detailed knowledge of those historical facts we can nevertheless make shrewd guesses on inductive grounds; these guesses are embodied in our intentional stance claims. Even if all this were true, it would tell us very little about the way such propensities were regulated by the internal machinery.

A currently more popular explanation is that the account of how the strategy works and the account of how the mechanism works will (roughly) *coincide:* for each predictively attributable belief, there will be a functionally salient internal state of the machinery, decomposable into functional parts in just about the same way the sentence expressing the belief is decomposable into parts—that is, words or terms. The inferences we attribute to rational creatures will be mirrored by physical, causal processes in the hardware; the *logical* form of the propositions believed will be copied in the *structural* form of the states in correspondence with them. This is the hypothesis that there is a *language of thought* coded in our brains, and our brains will eventually be understood as symbol manipulating systems in at least rough analogy with computers. Many different versions of this view are currently being explored, in the new research program called cognitive science, and provided one allows great latitude for attenuation of the basic, bold claim, I think some version of it will prove correct.

But I do not believe that this is *obvious.* Those who think that it is obvious, or inevitable, that such a theory will prove true (and there are many who do), are confusing two different empirical claims. The first is that intentional stance description yields an objective, real pattern in the world—the pattern our imaginary Martians missed. This is an empirical claim, but one that is confirmed beyond skepticism. The second is that this real pattern is *produced by* another real pattern roughly isomorphic to it within the brains of intelligent creatures. Doubting the existence of the second real pattern is not doubting the existence of the first. There *are* reasons for believing in the second pattern, but they are not overwhelming. The best simple account I can give of the reasons is as follows.

As we ascend the scale of complexity from simple thermostat, through sophisticated robot, to human being, we discover that our efforts to design systems with the requisite behavior increasingly run foul of the problem of *combinatorial explosion.* Increasing some parameter by, say, ten percent—ten percent more inputs or more degrees of freedom in the behavior to be controlled or more words to be recognized or whatever—tends to increase the internal complexity of the system being designed by orders of magnitude. Things get out of hand very fast and, for instance, can lead to computer programs that will swamp the largest, fastest machines. Now somehow the brain has

solved the problem of combinatorial explosion. It is a gigantic network of billions of cells, but still finite, compact, reliable, and swift, and capable of learning new behaviors, vocabularies, theories, almost without limit. Some elegant, *generative*, indefinitely extendable principles of representation must be responsible. We have only one model of such a representation system: a human language. So the argument for a language of thought comes down to this: what else could it be? We have so far been unable to imagine any plausible alternative in any detail. That is a good enough reason, I think, for recommending as a matter of scientific tactics that we pursue the hypothesis in its various forms as far as we can.[6] But we will engage in that exploration more circumspectly, and fruitfully, if we bear in mind that its inevitable rightness is far from assured. One does not well understand even a true empirical hypothesis so long as one is under the misapprehension that it is necessarily true.

BIBLIOGRAPHY

Akins, K. A., "On piranhas, narcissism, and mental representation," CCM-86-2, 1986. Center for Cognitive Studies, Tufts University.

Cherniak, C. *Minimal Rationality* (Cambridge, MA: MIT Press, 1986).

Churchland, P. M. *Scientific Realism and the Plasticity of Mind* (Cambridge, UK: Cambridge University Press, 1979).

Dennett, D. C., "Intentional systems," *Journal of Philosophy* 8, no. 87 (1971): p. 106. Reprinted in Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology* (Cambridge, MA: MIT Press, 1978).

_____ "Conditions of personhood," in *The Identities of Persons*, A. Rorty, ed., (Berkeley: University of California Press, 1976). Reprinted in Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology*.

_____ *Brainstorms: Philosophical Essays on Mind and Psychology*.

_____ *The Intentional Stance* (Cambridge, MA: MIT Press, 1987).

Feyerabend, P. *Science in a Free Society* (London: New Left Bank Publishers, 1978).

Hayes, P., "The naïve physics manifesto," in *Expert Systems in the Microelectronic Age*, D. Michie, ed., (Edinburgh: Edinburgh University Press, 1979).

McCarthy, J., "Ascribing mental qualities to machines," in *Philosophical Perspectives on Artificial Intelligence*, M. Ringle, ed., (Atlantic Highlands, NJ: Humanities Press, 1979).

NOTES

1. The idea that most of anyone's beliefs *must* be true seems obvious to some people. Support for the idea can be found in works by Quine, Putnam, Shoemaker, Davidson, and myself. Other people find the idea equally incredible—so probably each side is calling a different phenomenon belief. Once one makes the distinction between belief and opinion (in my technical sense—see "How to Change Your Mind" in Dennett 1978, chapter 16), according to which opinions are linguistically infected, relatively sophisticated cognitive states—*roughly* states of betting on the truth of a particular, formulated sentence—one can see the near triviality of the claim that most beliefs are true. A few reflections on peripheral matters should bring it out. Consider Democritus, who had a systematic, all-embracing, but (let us say, for the sake of argument) entirely false physics. He had things *all wrong*, though his views held together and had a sort of systematic utility. But even if every *claim* that scholarship permits us to attribute to Democritus (either explicit or implicit in his writings) is false, these represent a vanishingly small fraction of his *beliefs*, which include both the vast numbers of humdrum standing beliefs he must have had (about which house he lived in, what to look for in a good pair of sandals, and so forth) and also those occasional beliefs that came and went by the millions as his perceptual experience changed.

   But, it may be urged, this isolation of his humdrum beliefs from his science relies on an insupportable distinction between truths of observation and truths of theory; all Democritus's beliefs are theory-laden, and since his theory is false, they are false. The reply is as follows: Granted that all observation beliefs are theory laden, why should we choose Democritus's *explicit*, sophisticated theory (couched in his *opinions*) as the theory with which to burden his quotidian observations? Note that the least theoretical compatriot of Democritus also had myriads of theory-laden observation beliefs—and was, in one sense, none the wiser for it. Why should we not suppose Democritus's observations are laden with the same (presumably innocuous) theory? If Democritus forgot his theory, or changed his mind, his observational beliefs would be *largely* untouched. To the extent that his sophisticated theory played a discernible role in his routine behavior and expectations and so forth, it would be quite appropriate to couch his humdrum beliefs in terms of the sophisticated theory, but this will not yield a *mainly false* catalogue of beliefs, since so few of his beliefs will be affected. (The effect of theory on observation is nevertheless often underrated. See Churchland 1979 for dramatic and convincing examples of the tight relationship that can sometimes exist between theory and experience. [The discussion in this note

was distilled from a useful conversation with Paul and Patricia Churchland and Michael Stack.])

2. A member of the audience in Oxford pointed out that if the Martian included the Earthling in his physical stance purview (a possibility I had not explicitly excluded), he would not be surprised by the Earthling's prediction. He would indeed have predicted exactly the pattern of X-ray modulations produced by the Earthling speaking Martian. True, but as the Martian wrote down the results of his calculations, his prediction of the Earthling's prediction would appear, word by Martian word, as on a Ouija board, and what would be baffling to the Martian was how this chunk of mechanism, the Earthling predictor dressed up like a Martian, was able to yield this *true* sentence of Martian when it was so informationally isolated from the events the Martian needed to know of in order to make his own prediction about the arriving automobile.
3. Might there not be intelligent beings who had no use for communicating, predicting, observing . . . ? There might be marvelous, nifty, invulnerable entities lacking these modes of action, but I cannot see what would lead us to call them *intelligent.*
4. John McCarthy's analogy to cryptography nicely makes this point. The larger the corpus of cipher text, the less chance there is of dual, systematically unrelated decipherings. For a very useful discussion of the principles and presuppositions of the intentional stance applied to machines—explicitly including thermostats—see McCarthy 1979.
5. This idea is the ancestor in effect of the species of different ideas lumped together under the rubric of *de re* belief. If one builds from this idea toward its scions, one can see better the difficulties with them, and how to repair them. (For more on this topic, see Dennett 1987, chapter 5.)
6. The fact that all *language of thought* models of mental representation so far proposed fall victim to combinatorial explosion in one way or another should temper one's enthusiasm for engaging in what Fodor aptly calls 'the only game in town.'

Eliminative Materialism and the Propositional Attitudes

Paul M. Churchland

Eliminative materialism is the thesis that our common-sense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience. Our mutual understanding and even our introspection may then be reconstituted within the conceptual framework of completed neuroscience, a theory we may expect to be more powerful by far than the common-sense psychology it displaces, and more substantially integrated within physical science generally. My purpose in this paper is to explore these projections, especially as they bear on (1) the principal elements of common-sense psychology: the propositional attitudes (beliefs, desires, etc.), and (2) the conception of rationality in which these elements figure.

This focus represents a change in the fortunes of materialism. Twenty years ago, emotions, qualia, and 'raw feels' were held to be the principal stumbling blocks for the materialist program. With these barriers dissolving,[1] the locus of opposition has shifted. Now it is the realm of the intentional, the realm of the propositional attitude, that is most commonly held up as being both irreducible to and ineliminable in favor of anything from within a materialist framework. Whether and why this is so, we must examine.

Such an examination will make little sense, however, unless it is first appreciated that the relevant network of common-sense concepts does indeed constitute an empirical theory, with all the functions, virtues, *and perils* entailed by that status. I shall therefore begin with a brief sketch of this view and a summary rehearsal of its rationale. The resistance it encounters still surprises me. After all, common sense has yielded up many theories. Recall the view that space has a preferred direction in which all things fall; that weight is an intrinsic feature of a body; that a force-free moving object will promptly return to rest; that the sphere of the

From *Journal of Philosophy* 78 (1981): pp. 67–90. Reprinted with permission of author and publisher.

heavens turns daily; and so on. These examples are clear, perhaps, but people seem willing to concede a theoretical component within common sense only if (1) the theory and the common sense involved are safely located in antiquity, and (2) the relevant theory is now so clearly false that its speculative nature is inescapable. Theories are indeed easier to discern under these circumstances. But the vision of hindsight is always 20/20. Let us aspire to some foresight for a change.

I. Why Folk Psychology Is a Theory

Seeing our common-sense conceptual framework for mental phenomena as a theory brings a simple and unifying organization to most of the major topics in the philosophy of mind, including the explanation and prediction of behavior, the semantics of mental predicates, action theory, the other-minds problem, the intentionality of mental states, the nature of introspection, and the mind-body problem. Any view that can pull this lot together deserves careful consideration.

Let us begin with the explanation of human (and animal) behavior. The fact is that the average person is able to explain, and even predict, the behavior of other persons with a facility and success that is remarkable. Such explanations and predictions standardly make reference to the desires, beliefs, fears, intentions, perceptions, and so forth, to which the agents are presumed subject. But explanations presuppose laws—rough and ready ones, at least—that connect the explanatory conditions with the behavior explained. The same is true for the making of predictions, and for the justification of subjunctive and counterfactual conditionals concerning behavior. Reassuringly, a rich network of common-sense laws can indeed be reconstructed from this quotidian commerce of explanation and anticipation; its principles are familiar homilies; and their sundry functions are transparent. Each of us understands others, as well as we do, because we share a tacit command of an integrated body of lore concerning the law-like relations holding among external circumstances, internal states, and overt behavior. Given its nature and functions, this body of lore may quite aptly be called 'folk psychology.'[2]

This approach entails that the semantics of the terms in our familiar mentalistic vocabulary is to be understood in the same manner as the semantics of theoretical terms generally: the meaning of any theoretical term is fixed or constituted by the network of laws in which it figures. (This position is quite distinct from logical behaviorism. We deny that the relevant laws are analytic, and it is the lawlike connections generally that carry the semantic weight, not just the connections with overt behavior. But this view does account for what little plausibility logical behaviorism did enjoy.)

More importantly, the recognition that folk psychology is a theory provides a simple and decisive solution to an old skeptical problem, the problem of other minds. The problematic conviction that another individual is the subject of certain mental states is not inferred deductively from his behavior, nor is it inferred by inductive analogy from the perilously isolated instance of one's own case. Rather, that conviction is a singular *explanatory hypothesis* of a perfectly straightforward kind. Its function, in conjunction with the background laws of folk psychology, is to provide explanations/predictions/understanding of the individual's continuing behavior, and it is credible to the degree that it is successful in this regard over competing hypotheses. In the main, such hypotheses are successful, and so the belief that others enjoy the internal states comprehended by folk psychology is a reasonable belief.

Knowledge of other minds thus has no essential dependence on knowledge of one's own mind. Applying the principles of our folk psychology to our behavior, a Martian could justly ascribe to us the familiar run of mental states, even though his own psychology were very different from ours. He would not, therefore, be 'generalizing from his own case.'

As well, introspective judgments about one's own case turn out not to have any special status or integrity anyway. On the present view, an introspective judgment is just an instance of an acquired habit of conceptual response to one's internal states, and the integrity of any particular response is always contingent on the integrity of the acquired conceptual framework (theory) in which the response is framed. Accordingly, one's *introspective* certainty that one's mind is the seat of beliefs and desires may be as badly misplaced as was the classical man's *visual* certainty that the star-flecked sphere of the heavens turns daily.

Another conundrum is the intentionality of mental states. The 'propositional attitudes,' as Russell called them, form the systematic core of folk psychology; and their uniqueness and

anomalous logical properties have inspired some to see here a fundamental contrast with anything that mere physical phenomena might conceivably display. The key to this matter lies again in the theoretical nature of folk psychology. The intentionality of mental states here emerges not as a mystery of nature, but as a structural feature of the concepts of folk psychology. Ironically, those same structural features reveal the very close affinity that folk psychology bears to theories in the physical sciences. Let me try to explain.

Consider the large variety of what might be called 'numerical attitudes' appearing in the conceptual framework of physical science: '. . . has a mass_{kg} of *n*,' '. . . has a velocity of *n*,' '. . . has a temperature_{K} of *n*,' and so forth. These expressions are predicate-forming expressions: when one substitutes a singular term for a number into the place held by '*n*,' a determinate predicate results. More interestingly, the relations between the various 'numerical attitudes' that result are precisely the relations between the numbers 'contained' in those attitudes. More interesting still, the argument place that takes the singular terms for numbers is open to quantification. All this permits the expression of generalizations concerning the lawlike relations that hold between the various numerical attitudes in nature. Such laws involve quantification over numbers, and they exploit the mathematical relations holding in that domain. Thus, for example,

(1) $(x)\,(f)\,(m)[((x \text{ has a mass of } m)\ \&\ (x \text{ suffers a net force of } f)) \supset (x \text{ accelerates at } f/m)]$

Consider now the large variety of propositional attitudes: '. . . believes that *p*,' '. . . desires that *p*,' '. . . fears that *p*,' '. . . is happy that *p*,' etc. These expressions are predicate-forming expressions also. When one substitutes a singular term for a proposition into the place held by '*p*,' a determinate predicate results, e.g., '. . . believes that Tom is tall.' (Sentences do not generally function as singular terms, but it is difficult to escape the idea that when a sentence occurs in the place held by '*p*,' it is there functioning as or like a singular term. On this, more below.) More interestingly, the relations between the resulting propositional attitudes are characteristically the relations that hold between the propositions 'contained' in them, relations such as entailment, equivalence, and mutual inconsistency. More interesting still, the argument place that takes the singular terms for propositions is open to quantification. All this permits the expression of generalizations concerning the lawlike relations that hold among propositional attitudes. Such laws involve quantification over propositions, and they exploit various relations holding in that domain. Thus, for example,

(2) $(x)\,(p)[(x \text{ fears that } p) \supset (x \text{ desires that } \sim p)]$

(3) $(x)\,(p)[((x \text{ hopes that } p)\ \&\ (x \text{ discovers that } p)) \supset (x \text{ is pleased that } p)]$

(4) $(x)\,(p)\,(q)[((x \text{ believes that } p)\ \&\ (x \text{ believes that (if } p \text{ then } q))) \supset (\text{barring confusion, distraction, etc., } x \text{ believes that } q)]$

(5) $(x)\,(p)\,(q)[((x \text{ desires that } p)\ \&\ (x\text{believes that (if } q \text{ then } p))\ \&\ (x \text{ is able to bring it about that } q)) \supset (\text{barring conflicting desires or preferred strategies, } x \text{ brings it about that } q)]$[3]

Not only is folk psychology a theory, it is so *obviously* a theory that it must be held a major mystery why it has taken until the last half of the twentieth century for philosophers to realize it. The structural features of folk psychology parallel perfectly those of mathematical physics; the only difference lies in the respective domain of abstract entities they exploit—numbers in the case of physics, and propositions in the case of psychology.

Finally, the realization that folk psychology is a theory puts a new light on the mind–body problem. The issue becomes a matter of how the ontology of one theory (folk psychology) is, or is not, going to be related to the ontology of another theory (completed neuroscience); and the major philosophical positions on the mind–body problem emerge as so many different anticipations of what future research will reveal about the intertheoretic status and integrity of folk psychology.

The identity theorist optimistically expects that folk psychology will be smoothly *reduced* by completed neuroscience, and its ontology preserved by dint of transtheoretic identities. The dualist expects that it will prove *ir*reducible to completed neuroscience, by dint of being a nonredundant description of an autonomous, nonphysical domain of natural phenomena. The functionalist also expects that it will prove irreducible, but on the quite different grounds that the internal economy characterized by folk psychology is not, in the last analysis, a law-governed economy of natural states, but an abstract organization of functional states, an organization instantiable in a variety of quite

different material substrates. It is therefore irreducible to the principles peculiar to any of them.

Finally, the eliminative materialist is also pessimistic about the prospects for reduction, but his reason is that folk psychology is a radically inadequate account of our internal activities, too confused and too defective to win survival through intertheoretic reduction. On his view it will simply be displaced by a better theory of those activities.

Which of these fates is the real destiny of folk psychology, we shall attempt to divine presently. For now, the point to keep in mind is that we shall be exploring the fate of a theory, a systematic, corrigible, speculative *theory*.

II. Why Folk Psychology Might (Really) Be False

Given that folk psychology is an empirical theory, it is at least an abstract possibility that its principles are radically false and that its ontology is an illusion. With the exception of eliminative materialism, however, none of the major positions takes this possibility seriously. None of them doubts the basic integrity or truth of folk psychology (hereafter, 'FP'), and all of them anticipate a future in which its laws and categories are conserved. This conservatism is not without some foundation. After all, FP does enjoy a substantial amount of explanatory and predictive success. And what better grounds than this for confidence in the integrity of its categories?

What better grounds indeed? Even so, the presumption in FP's favor is spurious, born of innocence and tunnel vision. A more searching examination reveals a different picture. First, we must reckon not only with FP's successes, but with its explanatory failures, and with their extent and seriousness. Second, we must consider the long-term history of FP, its growth, fertility, and current promise of future development. And third, we must consider what sorts of theories are *likely* to be true of the etiology of our behavior, given what else we have learned about ourselves in recent history. That is, we must evaluate FP with regard to its coherence and continuity with fertile and well-established theories in adjacent and overlapping domains—with evolutionary theory, biology, and neuroscience, for example—because active coherence with the rest of what we presume to know is perhaps the final measure of any hypothesis.

A serious inventory of this sort reveals a very troubled situation, one which would evoke open skepticism in the case of any theory less familiar and dear to us. Let me sketch some relevant detail. When one centers one's attention not on what FP can explain, but on what it cannot explain or fails even to address, one discovers that there is a very great deal. As examples of central and important mental phenomena that remain largely or wholly mysterious within the framework of FP, consider the nature and dynamics of mental illness, the faculty of creative imagination, or the ground of intelligence differences between individuals. Consider our utter ignorance of the nature and psychological functions of sleep, that curious state in which a third of one's life is spent. Reflect on the common ability to catch an outfield fly ball on the run, or hit a moving car with a snowball. Consider the internal construction of a 3-D visual image from subtle differences in the 2-D array of stimulations in our respective retinas. Consider the rich variety of perceptual illusions, visual and otherwise. Or consider the miracle of memory, with its lightning capacity for relevant retrieval. On these and many other mental phenomena, FP sheds negligible light.

One particularly outstanding mystery is the nature of the learning process itself, especially where it involves large-scale conceptual change, and especially as it appears in its pre-linguistic or entirely nonlinguistic form (as in infants and animals), which is by far the most common form in nature. FP is faced with special difficulties here, since its conception of learning as the manipulation and storage of propositional attitudes founders on the fact that how to formulate, manipulate, and store a rich fabric of propositional attitudes is itself something that is learned, and is only one among many acquired cognitive skills. FP would thus appear constitutionally incapable of even addressing this most basic of mysteries.[4]

Failures on such a large scale do not (yet) show that FP is a false theory, but they do move that prospect well into the range of real possibility, and they do show decisively that FP is *at best* a highly superficial theory, a partial and unpenetrating gloss on a deeper and more complex reality. Having reached this opinion, we may be forgiven for exploring the possibility that FP provides a positively misleading sketch of our internal kinematics and dynamics, one whose success is owed more to selective application and forced interpretation on our part than to genuine theoretical insight on FP's part.

A look at the history of FP does little to allay such fears, once raised. The story is one of retreat, infertility, and decadence. The presumed

domain of FP used to be much larger than it is now. In primitive cultures, the behavior of most of the elements of nature were understood in intentional terms. The wind could know anger, the moon jealousy, the river generosity, the sea fury, and so forth. These were not metaphors. Sacrifices were made and auguries undertaken to placate or divine the changing passions of the gods. Despite its sterility, this animistic approach to nature has dominated our history, and it is only in the last two or three thousand years that we have restricted FP's literal application to the domain of the higher animals.

Even in this preferred domain, however, both the content and the success of FP have not advanced sensibly in two or three thousand years. The FP of the Greeks is essentially the FP we use today, and we are negligibly better at explaining human behavior in its terms than was Sophocles. This is a very long period of stagnation and infertility for any theory to display, especially when faced with such an enormous backlog of anomalies and mysteries in its own explanatory domain. Perfect theories, perhaps, have no need to evolve. But FP is profoundly imperfect. Its failure to develop its resources and extend its range of success is therefore darkly curious, and one must query the integrity of its basic categories. To use Imre Lakatos' terms, FP is a stagnant or degenerating research program, and has been for millennia.

Explanatory success to date is of course not the only dimension in which a theory can display virtue or promise. A troubled or stagnant theory may merit patience and solicitude on other grounds; for example, on grounds that it is the only theory or theoretical approach that fits well with other theories about adjacent subject matters, or the only one that promises to reduce to or be explained by some established background theory whose domain encompasses the domain of the theory at issue. In sum, it may rate credence because it holds promise of theoretical integration. How does FP rate in this dimension?

It is just here, perhaps, that FP fares poorest of all. If we approach *homo sapiens* from the perspective of natural history and the physical sciences, we can tell a coherent story of his constitution, development, and behavioral capacities which encompasses particle physics, atomic and molecular theory, organic chemistry, evolutionary theory, biology, physiology, and materialistic neuroscience. That story, though still radically incomplete, is already extremely powerful, outperforming FP at many points even in its own domain. And it is deliberately and self-consciously coherent with the rest of our developing world picture. In short, the greatest theoretical synthesis in the history of the human race is currently in our hands, and parts of it already provide searching descriptions and explanations of human sensory input, neural activity, and motor control.

But FP is no part of this growing synthesis. Its intentional categories stand magnificently alone, without visible prospect of reduction to that larger corpus. A successful reduction cannot be ruled out, in my view, but FP's explanatory impotence and long stagnation inspire little faith that its categories will find themselves neatly reflected in the framework of neuroscience. On the contrary, one is reminded of how alchemy must have looked as elemental chemistry was taking form, how Aristotelean cosmology must have looked as classical mechanics was being articulated, or how the vitalist conception of life must have looked as organic chemistry marched forward.

In sketching a fair summary of this situation, we must make a special effort to abstract from the fact that FP is a central part of our current *lebenswelt*, and serves as the principal vehicle of our interpersonal commerce. For these facts provide FP with a conceptual inertia that goes far beyond its purely theoretical virtues. Restricting ourselves to this latter dimension, what we must say is that FP suffers explanatory failures on an epic scale, that it has been stagnant for at least twenty-five centuries, and that its categories appear (so far) to be incommensurable with or orthogonal to the categories of the background physical science whose long-term claim to explain human behavior seems undeniable. Any theory that meets this description must be allowed a serious candidate for outright elimination.

We can of course insist on no stronger conclusion at this stage. Nor is it my concern to do so. We are here exploring a possibility, and the facts demand no more, and no less, than it be taken seriously. The distinguishing feature of the eliminative materialist is that he takes it very seriously indeed.

III. Arguments against Elimination

Thus the basic rationale of eliminative materialism: FP is a theory, and quite probably a false one; let us attempt, therefore to transcend it.

The rationale is clear and simple, but many find it uncompelling. It will be objected that FP

is not, strictly speaking, an *empirical* theory; that it is not false, or at least not refutable by empirical considerations; and that it ought not or cannot be transcended in the fashion of a defunct empirical theory. In what follows we shall examine these objections as they flow from the most popular and best-founded of the competing positions in the philosophy of mind: functionalism.

An antipathy toward eliminative materialism arises from two distinct threads running through contemporary functionalism. The first thread concerns the *normative* character of FP, or at least of that central core of FP which treats of the propositional attitudes. FP, some will say, is a characterization of an ideal, or at least praiseworthy mode of internal activity. It outlines not only what it is to have and process beliefs and desires, but also (and inevitably) what it is to be rational in their administration. The ideal laid down by FP may be imperfectly achieved by empirical humans, but this does not impugn FP as a normative characterization. Nor need such failures seriously impugn FP even as a descriptive characterization, for it remains true that our activities can be both usefully and accurately understood as rational *except for* the occasional lapse due to noise, interference, or other breakdown, which defects empirical research may eventually unravel. Accordingly, though neuroscience may usefully augment it, FP has no pressing need to be displaced, even as a descriptive theory; nor could it be replaced, qua normative characterization, by any descriptive theory of neural mechanisms, since rationality is defined over propositional attitudes like beliefs and desires. FP, therefore, is here to stay.

Daniel Dennett has defended a view along these lines.[5] And the view just outlined gives voice to a theme of the property dualists as well. Karl Popper and Joseph Margolis both cite the normative nature of mental and linguistic activity as a bar to their penetration or elimination by any descriptive/materialist theory.[6] I hope to deflate the appeal of such moves below.

The second thread concerns the *abstract* nature of FP. The central claim of functionalism is that the principles of FP characterize our internal states in a fashion that makes no reference to their intrinsic nature or physical constitution. Rather, they are characterized in terms of the network of causal relations they bear to one another, and to sensory circumstances and overt behavior. Given its abstract specification, that internal economy may therefore be realized in a nomically heterogeneous variety of physical systems. All of them may differ, even radically, in their physical constitution, and yet at another level, they will all share the same nature. This view, says Fodor, 'is compatible with very strong claims about the ineliminability of mental language from behavioral theories.'[7] Given the real possibility of multiple instantiations in heterogeneous physical substrates, we cannot eliminate the functional characterization in favor of any theory peculiar to one such substrate. That would preclude our being able to describe the (abstract) organization that any one instantiation shares with all the others. A functional characterization of our internal states is therefore here to stay.

This second theme, like the first, assigns a faintly stipulative character to FP, as if the onus were on the empirical systems to instantiate faithfully the organization that FP specifies, instead of the onus being on FP to describe faithfully the internal activities of a naturally distinct class of empirical systems. This impression is enhanced by the standard examples used to illustrate the claims of functionalism—mousetraps, valve-lifters, arithmetical calculators, computers, robots, and the like. These are artifacts, constructed to fill a preconceived bill. In such cases, a failure of fit between the physical system and the relevant functional characterization impugns only the former, not the latter. The functional characterization is thus removed from empirical criticism in a way that is most unlike the case of an empirical theory. One prominent functionalist—Hilary Putnam—has argued outright that FP is not a corrigible theory at all.[8] Plainly, if FP is construed on these models, as regularly it is, the question of its empirical integrity is unlikely ever to pose itself, let alone receive a critical answer.

Although fair to some functionalists, the preceding is not entirely fair to Fodor. On his view the aim of psychology is to find the *best* functional characterization of ourselves, and what that is remains an empirical question. As well, his argument for the ineliminability of mental vocabulary from psychology does not pick out current FP in particular as ineliminable. It need claim only that *some* abstract functional characterization must be retained, some articulation or refinement of FP perhaps.

His estimate of eliminative materialism remains low, however. First, it is plain that Fodor thinks there is nothing fundamentally or interestingly wrong with FP. On the contrary, FP's central conception of cognitive activity—as consisting in the manipulation of propositional

attitudes—turns up as the central element in Fodor's own theory on the nature of thought (*The Language of Thought*, op. cit.). And second, there remains the point that, whatever tidying up FP may or may not require, it cannot be displaced by any naturalistic theory of our physical substrate, since it is the abstract functional features of his internal states that make a person, not the chemistry of his substrate.

All of this is appealing. But almost none of it, I think, is right. Functionalism has too long enjoyed its reputation as a daring and *avant garde* position. It needs to be revealed for the short-sighted and reactionary position it is.

IV. The Conservative Nature of Functionalism

A valuable perspective on functionalism can be gained from the following story. To begin with, recall the alchemists' theory of inanimate matter. We have here a long and variegated tradition, of course, not a single theory, but our purposes will be served by a gloss.

The alchemists conceived the 'inanimate' as entirely continuous with animated matter, in that the sensible and behavioral properties of the various substances are owed to the ensoulment of baser matter by various spirits or essences. These nonmaterial aspects were held to undergo development, just as we find growth and development in the various souls of plants, animals, and humans. The alchemist's peculiar skill lay in knowing how to seed, nourish, and bring to maturity the desired spirits enmattered in the appropriate combinations.

On one orthodoxy, the four fundamental spirits (for 'inanimate' matter) were named 'mercury,' 'sulphur,' 'yellow arsenic,' and 'sal ammoniac.' Each of these spirits was held responsible for a rough but characteristic syndrome of sensible, combinatorial, and causal properties. The spirit mercury, for example, was held responsible for certain features typical of metallic substances—their shininess, liquefiability, and so forth. Sulphur was held responsible for certain residual features typical of metals, and for those displayed by the ores from which running metal could be distilled. Any given metallic substance was a critical orchestration principally of these two spirits. A similar story held for the other two spirits, and among the four of them a certain domain of physical features and transformations was rendered intelligible and controllable.

The degree of control was always limited, of course. Or better, such prediction and control as the alchemists possessed was owed more to the manipulative lore acquired as an apprentice to a master, than to any genuine insight supplied by the theory. The theory followed, more than it dictated, practice. But the theory did supply some rhyme to the practice, and in the absence of a developed alternative it was sufficiently compelling to sustain a long and stubborn tradition.

The tradition had become faded and fragmented by the time the elemental chemistry of Lavoisier and Dalton arose to replace it for good. But let us suppose that it had hung on a little longer—perhaps because the four-spirit orthodoxy had become a thumb-worn part of everyman's common sense—and let us examine the nature of the conflict between the two theories and some possible avenues of resolution.

No doubt the simplest line of resolution, and the one which historically took place, is outright displacement. The dualistic interpretation of the four essences—as immaterial spirits—will appear both feckless and unnecessary given the power of the corpuscularian taxonomy of atomic chemistry. And a reduction of the old taxonomy to the new will appear impossible, given the extent to which the comparatively toothless old theory cross-classifies things relative to the new. Elimination would thus appear the only alternative—*unless* some cunning and determined defender of the alchemical vision has the wit to suggest the following defense.

Being 'ensouled by mercury,' or 'sulphur,' or either of the other two so-called spirits, is actually a *functional* state. The first, for example, is defined by the disposition to reflect light, to liquefy under heat, to unite with other matter in the same state, and so forth. And each of these four states is related to the others, in that the syndrome for each varies as a function of which of the other three states is also instantiated in the same substrate. Thus the level of description comprehended by the alchemical vocabulary is abstract: various material substances, suitably 'ensouled,' can display the features of a metal, for example, or even of gold specifically. For it is the total syndrome of occurrent and causal properties which matters, not the corpuscularian details of the substrate. Alchemy, it is concluded, comprehends a level of organization in reality distinct from and irreducible to the organization found at the level of corpuscularian chemistry.

This view might have had considerable appeal. After all, it spares alchemists the burden

of defending immaterial souls that come and go; it frees them from having to meet the very strong demands of a naturalistic reduction; and it spares them the shock and confusion of outright elimination. Alchemical theory emerges as basically all right! Nor need they appear too obviously stubborn or dogmatic in this. Alchemy as it stands, they concede, may need substantial tidying up, and experience must be our guide. But we need not fear its naturalistic displacement, they remind us, since it is the particular orchestration of the syndromes of occurrent and causal properties which makes a piece of matter gold, not the idiosyncratic details of its corpuscularian substrate. A further circumstance would have made this claim even more plausible. For the fact is, the alchemists *did* know how to make gold, in this relevantly weakened sense of 'gold,' and they could do so in a variety of ways. Their 'gold' was never as perfect, alas, as the 'gold' nurtured in nature's womb, but what mortal can expect to match the skills of nature herself?

What this story shows is that it is at least possible for the constellation of moves, claims, and defenses characteristic of functionalism to constitute an outrage against reason and truth, and to do so with a plausibility that is frightening. Alchemy is a terrible theory, well-deserving of its complete elimination, and the defense of it just explored is reactionary, obfuscatory, retrograde, and wrong. But in historical context, that defense might have seemed wholly sensible, even to reasonable people.

The alchemical example is a deliberately transparent case of what might well be called 'the functionalist stratagem,' and other cases are easy to imagine. A cracking good defense of the phlogiston theory of combustion can also be constructed along these lines. Construe being highly phlogisticated and being dephlogisticated as functional states defined by certain syndromes of causal dispositions; point to the great variety of natural substrates capable of combustion and calxification; claim an irreducible functional integrity for what has proved to lack any natural integrity; and bury the remaining defects under a pledge to contrive improvements. A similar recipe will provide new life for the four humors of medieval medicine, for the vital essence or archeus of pre-modern biology, and so forth.

If its application in these other cases is any guide, the functionalist stratagem is a smokescreen for the preservation of error and confusion. Whence derives our assurance that in contemporary journals the same charade is not being played out on behalf of FP? The parallel with the case of alchemy is in all other respects distressingly complete, right down to the parallel between the search for artificial gold and the search for artificial intelligence!

Let me not be misunderstood on this last point. Both aims are worthy aims: thanks to nuclear physics, artificial (but real) gold is finally within our means, if only in submicroscopic quantities; and artificial (but real) intelligence eventually will be. But just as the careful orchestration of superficial syndromes was the wrong way to produce genuine gold, so may the careful orchestration of superficial syndromes be the wrong way to produce genuine intelligence. Just as with gold, what may be required is that our science penetrate to the underlying *natural* kind that gives rise to the total syndrome directly.

In summary, when confronted with the explanatory impotence, stagnant history, and systematic isolation of the intentional idioms of FP, it is not an adequate or responsive defense to insist that those idioms are abstract, functional, and irreducible in character. For one thing, this same defense could have been mounted with comparable plausibility no matter *what* haywire network of internal states our folklore had ascribed to us. And for another, the defense assumes essentially what is at issue: it assumes that it is the intentional idioms of FP, plus or minus a bit, that express the *important* features shared by all cognitive systems. But they may not. Certainly it is wrong to assume that they do, and then argue against the possibility of a materialistic displacement on grounds that it must describe matters at a level that is different from the important level. This just begs the question in favor of the older framework.

Finally, it is very important to point out that eliminative materialism is strictly *consistent* with the claim that the essence of a cognitive system resides in the abstract functional organization of its internal states. The eliminative materialist is not committed to the idea that the correct account of cognition *must* be a naturalistic account, though he may be forgiven for exploring the possibility. What he does hold is that the correct account of cognition, whether functionalistic or naturalistic, will bear about as much resemblance to FP as modern chemistry bears to four-spirit alchemy.

Let us now try to deal with the argument, against eliminative materialism, from the normative dimension of FP. This can be dealt with rather swiftly, I believe.

First, the fact that the regularities ascribed by the intentional core of FP are predicated on certain logical relations among propositions is not by itself grounds for claiming anything essentially normative about FP. To draw a relevant parallel, the fact that the regularities ascribed by the classical gas law are predicated on arithmetical relations between numbers does not imply anything essentially normative about the classical gas law. And logical relations between propositions are as much an objective matter of abstract fact as are arithmetical relations between numbers. In this respect, the law

(4) $(x)\,(p)\,(q)$ [((x believes that p) & (x believes that (if p then q))) ⊃ (barring confusion, distraction, etc., x believes that q)]

is entirely on a par with the classical gas law

(6) $(x)\,(P)\,(V)\,(\mu)$ [((x has a pressure P) & (x has a volume V) & (x has a quantity μ)) ⊃ (barring very high pressure or density, x has a temperature of $PV/\mu R$)]

A normative dimension enters only because we happen to *value* most of the patterns ascribed by FP. But we do not value all of them. Consider

(7) $(x)\,(p)$ [((x desires with all his heart that p) & (x learns that $\sim p$)) ⊃ (barring unusual strength of character, x is shattered that $\sim p$)]

Moreover, and as with normative convictions generally, fresh insight may motivate major changes in what we value.

Second, the laws of FP ascribe to us only a very minimal and truncated rationality, not an ideal rationality as some have suggested. The rationality characterized by the set of all FP laws falls well short of an ideal rationality. This is not surprising. We have no clear or finished conception of ideal rationality anyway; certainly the ordinary man does not. Accordingly, it is just not plausible to suppose that the explanatory failures from which FP suffers are owed primarily to human failure to live up to the ideal standard it provides. Quite to the contrary, the conception of rationality it provides appears limping and superficial, especially when compared with the dialectical complexity of our scientific history, or with the ratiocinative virtuosity displayed by any child.

Third, even if our current conception of rationality—and more generally, of cognitive virtue—is largely constituted within the sentential/propositional framework of FP, there is no guarantee that this framework is adequate to the deeper and more accurate account of cognitive virtue which is clearly needed. Even if we concede the categorial integrity of FP, at least as applied to language-using humans, it remains far from clear that the basic parameters of intellectual virtue are to be found at the categorial level comprehended by the propositional attitudes. After all, language use is something that is learned, by a brain already capable of vigorous cognitive activity; language use is acquired as only one among a great variety of learned manipulative skills; and it is mastered by a brain that evolution has shaped for a great many functions, language use being only the very latest and perhaps the least of them. Against the background of these facts, language use appears as an extremely peripheral activity, as a racially idiosyncratic mode of social interaction which is mastered thanks to the versatility and power of a more basic mode of activity. Why accept then, a theory of cognitive activity that models its elements on the elements of human language? And why assume that the fundamental parameters of intellectual virtue are or can be defined over the elements at this superficial level?

A serious advance in our appreciation of cognitive virtue would thus seem to *require* that we go beyond FP, that we transcend the poverty of FP's conception of rationality by transcending its propositional kinematics entirely, by developing a deeper and more general kinematics of cognitive activity, and by distinguishing within this new framework which of the kinematically possible modes of activity are to be valued and encouraged (as more efficient, reliable, productive, or whatever). Eliminative materialism thus does not imply the end of our normative concerns. It implies only that they will have to be reconstituted at a more revealing level of understanding, the level that a matured neuroscience will provide.

What a theoretically informed future might hold in store for us, we shall now turn to explore. Not because we can foresee matters with any special clarity, but because it is important to try to break the grip on our imagination held by the propositional kinematics of FP. As far as the present section is concerned, we may summarize our conclusions as follows. FP is nothing more and nothing less than a culturally entrenched theory of how we and the higher animals work. It has no special features that make it empirically invulnerable, no unique functions

that make it irreplaceable, no special status of any kind whatsoever. We shall turn a skeptical ear then, to any special pleading on its behalf.

V. Beyond Folk Psychology

What might the elimination of FP actually involve—not just the comparatively straightforward idioms for sensation, but the entire apparatus of propositional attitudes? That depends heavily on what neuroscience might discover, and on our determination to capitalize on it. Here follow three scenarios in which the operative conception of cognitive activity is progressively divorced from the forms and categories that characterize natural language. If the reader will indulge the lack of actual substance, I shall try to sketch some plausible form.

First suppose that research into the structure and activity of the brain, both fine-grained and global, finally does yield a new kinematics and correlative dynamics for what is now thought of as cognitive activity. The theory is uniform for all terrestrial brains, not just human brains, and it makes suitable conceptual contact with both evolutionary biology and non-equilibrium thermodynamics. It ascribes to us, at any given time, a set or configuration of complex states, which are specified within the theory as figurative 'solids' within a four- or five-dimensional phase space. The laws of the theory govern the interaction, motion, and transformation of these 'solid' states within that space, and also their relations to whatever sensory and motor transducers the system possesses. As with celestial mechanics, the exact specification of the 'solids' involved and the exhaustive accounting of all dynamically relevant adjacent 'solids' is not practically possible, for many reasons, but here also it turns out that the obvious approximations we fall back on yield excellent explanations/predictions of internal change and external behavior, at least in the short term. Regarding long-term activity, the theory provides powerful and unified accounts of the learning process, the nature of mental illness, and variations in character and intelligence across the animal kingdom as well as across individual humans.

Moreover, it provides a straightforward account of 'knowledge,' as traditionally conceived. According to the new theory, any declarative sentence to which a speaker would give confident assent is merely a one-dimensional *projection*—through the compound lens of Wernicke's and Broca's areas onto the idiosyncratic surface of the speaker's language—a one-dimensional projection of a four- or five-dimensional 'solid' that is an element in his true kinematical state. (Recall the shadows on the wall of Plato's cave.) Being projections of that inner reality, such sentences do carry significant information regarding it and are thus fit to function as elements in a communication system. On the other hand, being *sub*dimensional projections, they reflect but a narrow part of the reality projected. They are therefore *un*fit to represent the deeper reality in all its kinematically, dynamically, and even normatively relevant respects. That is to say, a system of propositional attitudes, such as FP, must inevitably fail to capture what is going on here, though it may reflect just enough superficial structure to sustain an alchemy-like tradition among folk who lack any better theory. From the perspective of the newer theory, however, it is plain that there simply are no law-governed states of the kind FP postulates. The real laws governing our internal activities are defined over different and much more complex kinematical states and configurations, as are the normative criteria for developmental integrity and intellectual virtue.

A theoretical outcome of the kind just described may fairly be counted as a case of elimination of one theoretical ontology in favor of another, but the success here imagined for systematic neuroscience need not have any sensible effect on common practice. Old ways die hard, and in the absence of some practical necessity, they may not die at all. Even so, it is not inconceivable that some segment of the population, or all of it, should become intimately familiar with the vocabulary required to characterize our kinematical states, learn the laws governing their interactions and behavioral projections, acquire a facility in their first-person ascription, and displace the use of FP altogether, even in the marketplace. The demise of FP's ontology would then be complete.

We may now explore a second and rather more radical possibility. Everyone is familiar with Chomsky's thesis that the human mind or brain contains innately and uniquely the abstract structures for learning and using specifically human natural languages. A competing hypothesis is that our brain does indeed contain innate structures, but that those structures have as their original and still primary function the organization of perceptual experience, the administration of linguistic categories being an acquired and additional function for which evolution has only incidentally suited them.[9] This hypothesis has

the advantage of not requiring the evolutionary saltation that Chomsky's view would seem to require, and there are other advantages as well. But these matters need not concern us here. Suppose, for our purposes, that this competing view is true, and consider the following story.

Research into the neural structures that fund the organization and processing of perceptual information reveals that they are capable of administering a great variety of complex tasks, some of them showing a complexity far in excess of that shown by natural language. Natural languages, it turns out, exploit only a very elementary portion of the available machinery, the bulk of which serves far more complex activities beyond the ken of the propositional conceptions of FP. The detailed unravelling of what that machinery is and of the capacities it has makes it plain that a form of language far more sophisticated than 'natural' language, though decidedly 'alien' in its syntactic and semantic structures, could also be learned and used by our innate systems. Such a novel system of communication, it is quickly realized, could raise the efficiency of information exchange between brains by an order of magnitude, and would enhance epistemic evaluation by a comparable amount, since it would reflect the underlying structure of our cognitive activities in greater detail than does natural language.

Guided by our new understanding of those internal structures, we manage to construct a new system of verbal communication entirely distinct from natural language, with a new and more powerful combinatorial grammar over novel elements forming novel combinations with exotic properties. The compounded strings of this alternative system—call them 'übersatzen'—are not evaluated as true or false, nor are the relations between them remotely analogous to the relations of entailment, etc., that hold between sentences. They display a different organization and manifest different virtues.

Once constructed, this 'language' proves to be learnable; it has the power projected; and in two generations it has swept the planet. Everyone uses the new system. The syntactic forms and semantic categories of so-called 'natural' language disappear entirely. And with them disappear the propositional attitudes of FP, displaced by a more revealing scheme in which (of course) 'übersatzenal attitudes' play the leading role. FP again suffers elimination.

This second story, note, illustrates a theme with endless variations. There are possible as many different 'folk psychologies' as there are possible differently structured communication systems to serve as models for them.

A third and even stranger possibility can be outlined as follows. We know that there is considerable lateralization of function between the two cerebral hemispheres, and that the two hemispheres make use of the information they get from each other by way of the great cerebral commissure—the corpus callosum—a giant cable of neurons connecting them. Patients whose commissure has been surgically severed display a variety of behavioral deficits that indicate a loss of access by one hemisphere to information it used to get from the other. However, in people with callosal agenesis (a congenital defect in which the connecting cable is simply absent), there is little or no behavioral deficit, suggesting that the two hemispheres have learned to exploit the information carried in other less direct pathways connecting them through the subcortical regions. This suggests that, even in the normal case, a developing hemisphere *learns* to make use of the information the cerebral commissure deposits at its doorstep. What we have then, in the case of a normal human, is two physically distinct cognitive systems (both capable of independent function) responding in a systematic and learned fashion to exchanged information. And what is especially interesting about this case is the sheer amount of information exchanged. The cable of the commissure consists of $\approx$ 200 million neurons,[10] and even if we assume that each of these fibres is capable of one of only two possible states each second (a most conservative estimate), we are looking at a channel whose information capacity is $> 2 \times 10^8$ binary bits/second. Compare this to the < 500 bits/second capacity of spoken English.

Now, if two distinct hemispheres can learn to communicate on so impressive a scale, why shouldn't two distinct brains learn to do it also? This would require an artificial 'commissure' of some kind, but let us suppose that we can fashion a workable transducer for implantation at some site in the brain that research reveals to be suitable, a transducer to convert a symphony of neural activity into (say) microwaves radiated from an aerial in the forehead, and to perform the reverse function of converting received microwaves back into neural activation. Connecting it up need not be an insuperable problem. We simply trick the normal processes of dendritic arborization into growing their own myriad connections with the active microsurface of the transducer.

Once the channel is opened between two or more people, they can learn (*learn*) to exchange information and coordinate their behavior with the same intimacy and virtuosity displayed by your own cerebral hemispheres. Think what this might do for hockey teams, and ballet companies, and research teams! If the entire population were thus fitted out, spoken language of any kind might well disappear completely, a victim of the 'why crawl when you can fly?' principle. Libraries become filled not with books, but with long recordings of exemplary bouts of neural activity. These constitute a growing cultural heritage, an evolving 'Third World,' to use Karl Popper's terms. But they do not consist of sentences or arguments.

How will such people understand and conceive of other individuals? To this question I can only answer, 'In roughly the same fashion that your right hemisphere 'understands' and 'conceives of' your left hemisphere—intimately and efficiently, but not propositionally!'

These speculations, I hope, will evoke the required sense of untapped possibilities, and I shall in any case bring them to a close here. Their function is to make some inroads into the aura of inconceivability that commonly surrounds the idea that we might reject FP. The felt conceptual strain even finds expression in an argument to the effect that the thesis of eliminative materialism is incoherent since it denies the very conditions presupposed by the assumption that it is meaningful. I shall close with a brief discussion of this very popular move.

As I have received it, the reductio proceeds by pointing out that the statement of eliminative materialism is just a meaningless string of marks or noises, unless that string is the expression of a certain *belief*, and a certain *intention* to communicate, and a *knowledge* of the grammar of the language, and so forth. But if the statement of eliminative materialism is true, then there are no such states to express. The statement at issue would then be a meaningless string of marks or noises. It would therefore *not* be true. Therefore it is not true. Q.E.D.

The difficulty with any nonformal reductio is that the conclusion against the initial assumption is always no better than the material assumptions invoked to reach the incoherent conclusion. In this case the additional assumptions involve a certain theory of meaning, one that presupposes the integrity of FP. But formally speaking, one can as well infer, from the incoherent result, that this theory of meaning is what must be rejected. Given the independent critique of FP levelled earlier, this would even seem the preferred option. But in any case, one cannot simply assume that particular theory of meaning without begging the question at issue, namely, the integrity of FP.

The question-begging nature of this move is most graphically illustrated by the following analogue, which I owe to Patricia Churchland.[11] The issue here, placed in the seventeenth century, is whether there exists such a substance as *vital spirit.* At the time, this substance was held, without significant awareness of real alternatives, to be that which distinguished the animate from the inanimate. Given the monopoly enjoyed by this conception, given the degree to which it was integrated with many of our other conceptions, and given the magnitude of the revisions any serious alternative conception would require, the following refutation of any anti-vitalist claim would be found instantly plausible.

> The anti-vitalist says that there is no such thing as vital spirit. But this claim is self-refuting. The speaker can expect to be taken seriously only if his claim cannot. For if the claim is true, then the speaker does not have vital spirit and must be *dead.* But if he is dead, then his statement is a meaningless string of noises, devoid of reason and truth.

The question-begging nature of this argument does not, I assume, require elaboration. To those moved by the earlier argument, I commend the parallel for examination.

The thesis of this paper may be summarized as follows. The propositional attitudes of folk psychology do not constitute an unbreachable barrier to the advancing tide of neuroscience. On the contrary; the principled displacement of folk psychology is not only richly possible, it represents one of the most intriguing theoretical displacements we can currently imagine.

NOTES

An earlier draft of this paper was presented at the University of Ottawa and to the Brain, Mind, and Person colloquium at SUNY/Oswego. My thanks for the suggestions and criticisms that have informed the present version.

1. See Paul Feyerabend, "Materialism and the Mind-Body Problem," *Review* of *Metaphysics* XVII 1, 65 (September 1963): pp. 49–66; Richard Rorty, "Mind-Body Identity, Privacy, and Categories," ibid., XIX 1, 73 (September 1965): pp. 24–54; and my *Scientific Realism and the Plasticity of Mind* (New York: Cambridge University Press, 1979).
2. We shall examine a handful of these laws presently. For a more comprehensive sampling of the laws of folk psychology, see my *Scientific Realism*

and Plasticity of Mind, op. cit., chapter 4. For a detailed examination of the folk principles that underwrite action explanations in particular, see my "The Logical Character of Action Explanations," *Philosophical Review* LXXIX, no. 2 (April 1970): pp. 214–36.

3. Staying within an objectual interpretation of the quantifiers, perhaps the simplest way to make systematic sense of expressions like ⌜*x* believes that *p*⌝ and closed sentences formed therefrom is just to construe whatever occurs in the nested position held by '*p*,' '*q*,' etc. as there having the function of a singular term. Accordingly, the standard connectives, as they occur between terms in that nested position, must be construed as there functioning as operators that form compound singular terms from other singular terms, and not as sentence operators. The compound singular terms so formed denote the appropriate compound propositions. Substitutional quantification will of course underwrite a different interpretation, and there are other approaches as well. Especially appealing is the prosentential approach of Dorothy Grover, Joseph Camp, and Nuel Belnap, "A Prosentential Theory of Truth," *Philosophical Studies* XXVII, no. 2 (February 1975): pp. 73–125. But the resolution of these issues is not vital to the present discussion.
4. A possible response here is to insist that the cognitive activity of animals and infants is linguaformal in its elements, structures, and processing right from birth. J. A. Fodor, in *The Language of Thought* (New York: Crowell 1975), has erected a positive theory of thought on the assumption that the innate forms of cognitive activity have precisely the form here denied. For a critique of Fodor's view, see Patricia Churchland, "Fodor on Language Learning," *Synthese* XXXVIII, no. 1 (May 1978): pp. 149–59.
5. Most explicitly in "Three Kinds of Intentional Psychology" in *Reduction, Time, and Reality*, R. Healey, ed., (New York: Cambridge University Press, 1981), pp. 37–61, but this theme of Dennett's goes all the way back to his "Intentional Systems," *The Journal of Philosophy*, LXVIII, no. 4 (February 25, 1971): pp. 87–106; reprinted in his *Brainstorms* (Montgomery, VT: Bradford Books, 1978).
6. Karl Popper, *Objective Knowledge* (New York: Oxford University Press, 1972); with J. Eccles, *The Self and Its Brain* (New York: Springer Verlag, 1978). Joseph Margolis, *Persons and Minds* (Boston: Reidel, 1978).
7. *Psychological Explanation* (New York: Random House, 1968), p. 116.
8. "Robots: Machines or Artificially Created Life?" *The Journal of Philosophy* LXI, no. 21 (November 12, 1964): pp. 668–91, pp. 675, 681 ff.
9. Richard Gregory defends such a view in "The Grammar of Vision," *Listener* LXXXIII, no. 2133 (February 1970): pp. 242–46; reprinted in his *Concepts and Mechanisms of Perception* (London: Duckworth, 1975), pp. 622–29.
10. M. S. Gazzaniga and J. E. LeDoux, *The Integrated Mind* (New York: Plenum Press, 1975).
11. "Is Determinism Self-Refuting?" *Mind* 90 (1981): pp. 99–101.

Alief and Belief*

Tamar Gendler

In March 2007, 4,000 feet above the floor of the Grand Canyon, a horseshoe-shaped cantilevered glass walkway was opened to the public. Extending 70 feet from the canyon's rim, the Grand Canyon Skywalk soon drew hundreds of visitors each day, among them *New York Times* reporter Edward Rothstein, who filed the following dispatch:

> A visitor to these stark and imposing lands of the Hualapai Indians on the western rim of the Grand Canyon knows what sensation is being promised at the journey's climax. After driving for a half-hour over bone-jolting dirt roads . . . you take a shuttle bus from the parking lot . . . . You deposit all cameras at a security desk, slip on yellow surgical booties and stride out onto a horseshoe-shaped walkway with transparent sides and walls that extends 70 feet into space, seemingly unsupported.
>
> Below the floor's five layers of glass (protected from scratches by the booties) can be seen the cracked, sharp-edged rock face of the canyon's rim and a drop of thousands of feet to the chasm below. The promise is the dizzying thrill of vertigo.
>
> And indeed, last week some visitors to this steel-supported walkway anchored in rock felt precisely that. One woman, her left hand desperately grasping the 60-inch-high glass sides and the other clutching the arm of a patient security guard, didn't dare move toward the transparent center of the walkway. The words imprinted on

From *The Journal of Philosophy*, 105:10, 2008, pp. 634-63.

> the $20 souvenir photographs taken of many venturesome souls herald completion of a daredevil stunt: 'I did it!!!'[1]

Though some readers may find this story politically or aesthetically disturbing, none—I take it—find it perplexing.[2] While the sarcasm of 'venturesome souls' is surely well placed, and the price of the "I did it!!!" photo is surely excessive, the basic phenomenon—that stepping onto a high transparent safe surface can induce feelings of vertigo—is both familiar and unmysterious.[3]

How should we describe the cognitive state of those who manage to stride to the Skywalk's center? Surely they *believe* that the walkway will hold: no one would willingly step onto a mile-high platform if they had even a scintilla of doubt concerning its stability. But alongside that belief there is something else going on. Although the venture-some souls wholeheartedly *believe* that the walkway is completely safe, they also *alieve* something very different. The alief has roughly the following content: 'Really high up, long long way down. Not a safe place to be! Get off!!'[4]

•

In a series of ingenious studies spanning several decades, psychologist Paul Rozin has demonstrated a widespread tendency for well-educated Western adults to exhibit behaviors consonant with a commitment to the existence of 'laws of sympathetic magic:'[5] that 'there can be a permanent transfer of properties from one object . . . to another by brief contact' (*contagion*) and that 'the action taken on an object affects similar objects' (*similarity*).[6]

So, for example, subjects are reluctant to drink from a glass of juice in which a completely sterilized dead cockroach has been stirred, hesitant to wear a laundered shirt that has been previously worn by someone they dislike, and loath to eat soup from a brand-new bedpan. They are disinclined to put their mouths on a piece of newly purchased vomit-shaped rubber (though perfectly willing to do so with sink stopper of similar size and material), averse to eating fudge that has been formed into the shape of dog feces, and far less accurate in throwing darts at pictures of faces of people they like than at neutral faces.[7]

How should we describe the cognitive state of those who hesitate to eat the feces-shaped fudge or wear their adversary's shirt? Surely they *believe* that the fudge has not changed its chemical composition, and that the shirt does not bear cooties[8]—just as they believe that that the newly purchased bedpan is sterile and that the fake vomit is actually made of rubber: asked directly, subjects show no hesitation in endorsing such claims. But alongside these beliefs there is something else going on. Although they *believe* that the items in question are harmless, they also *alieve* something very different. The alief has roughly the following content: 'Filthy object! Contaminated! Stay away!'

•

Last month, when I was traveling to the APA Program Committee meeting, I accidentally left my wallet at home. I noticed its absence when I arrived at the check-in desk at the Hartford Airport and fully expected to be turned away from my flight. Much to my surprise, the desk agent simply wrote the words 'No ID' on my boarding pass and told me to allow for a few extra minutes at security.[9] The various scans showed nothing amiss, so I boarded my plane, flew to Baltimore, and made my way to the meeting site.

Though the TSA may not require identification, restaurants and hotels do require payment, so when I got to Baltimore, I arranged to borrow money from a friend who was also attending the meeting. As he handed me the bills, I said: 'Thanks so much for helping me out like this. It is really important for me to have this much cash since I don't have my wallet.' Rooting through my bag as I talked, I continued: 'It's a lot of cash to be carrying loose, though, so let me just stash it in my wallet. . . .'

How should we describe my mental state as my fingers searched for my wallet to house the explicitly wallet-compensatory money? Surely I *believed* that I had left my wallet in New Haven; after all, the reason I was borrowing so much money was because I knew I had no credit cards or cash with me. But alongside that belief there was something else going on. Although I *believed* that my wallet was several hundred miles away as I rooted through my bag, I simultaneously *alieved* something very different. The alief had roughly the following content: 'Bunch of money. Needs to go into a safe place. Activate wallet-retrieval motor routine now.'

•

> Charles is watching a horror movie about a terrible green slime. He cringes in his seat as the slime oozes slowly but relentlessly over the earth destroying everything in its path. Soon a greasy head emerges from the undulating mass, and two beady eyes roll around, finally fixing on the camera. The slime, picking up speed, oozes on a new course straight towards the viewers. Charles emits a shriek and clutches desperately at his chair.[10]

How should we describe Charles's cognitive state? Surely he does not *believe* that that he is in physical peril; as Kendall Walton writes 'Charles knows perfectly well that the slime is not real and that he is in no danger' (ibid., p. 6). But alongside that belief there is something else going on. Although Charles *believes* that he is sitting safely in a chair in a theater in front of a movie screen, he also *alieves* something very different. The alief has roughly the following content: 'Dangerous two-eyed creature heading towards me! H-e-l-p ! Activate fight or flight adrenaline now!'

I. Introducing Alief

I.1. Belief-Behavior Mismatch and Belief-Discordant Alief. In each of the cases presented above, it seems clear what the subject believes[11]: that the walkway is safe, that the substance is edible or potable, that the wallet is in New Haven, that the theater is in no danger of being invaded by slime, and so on. Ask the subject directly and she will show no hesitation in endorsing such claims as true. Ask her to bet, and this is where she will place her money. Ask her to think about what her other beliefs imply and this is what she will conclude. Look at her overarching behavior and this is what it will point to. At the same time, the belief fails to be accompanied by certain belief-appropriate behaviors and attitudes: something is awry.

When else do we find this sort of belief-behavior mismatch? One sort of case is that of deliberate deception. If I believe that I have a winning hand, but I am trying to mislead you into thinking that I do not, I will behave in ways discordant with my belief. But clearly, this is not a good model for the cases just considered: Charles is not trying to fool the movie-maker; Rozin's subjects are not trying to mislead the experimenters. In contrast to the cases of deliberate deception, the belief-behavior mismatch in our cases is not the result of something other-directed and deliberately controlled.

Perhaps, then, it is akin to a case of self-deception? A self-deceived subject believes, say, that her child has committed some terrible crime, but somehow brings herself to represent the situation—both to herself and to others—as if she believed precisely the opposite, resulting in the requisite belief-behavior mismatch.[12] This is an improvement on the previous model; it corrects the problem of other-directedness, and—to some extent—the problem of deliberate control. But it still misrepresents the structure of the situation: it is not that the reluctant walker on the Hulapai Skywalk believes that the surface is safe, but has somehow deceived herself into thinking that it is risky; it is not that Rozin's subject believes that the bedpan is sterile, but somehow deceives herself into thinking that there's some reason not to drink from it. The mismatch runs two directions: unlike in cases of self-deception, the subjects in our cases show no reluctance to endorse explicitly the belief with which their behavior fails to accord. And unlike in cases of self-deception, their behavioral responses do not result from some deliberate or quasi-deliberate process of misrepresentation.

Perhaps, then, the subjects' hesitation to act on their beliefs is the result of some sort of doubt or uncertainty? In setting out for the day, I might dither a bit before leaving my umbrella at home: 'it's not going to rain,' I might aver—though I am not completely certain that I am right. Though the action-pattern is strikingly similar to some of the cases above, the model is still inadequate. Stepping onto the Skyway, eating the stool-shaped fudge, or staying seated in the theater is not like willing oneself to play Russian roulette: it is not a case of discounting a low-probability outcome and hoping for the best. Charles does not leave the theater thinking: 'Phew! It's lucky the slime stayed on the screen this time!' Rozin's subject does not breathe a sigh of relief that the dart hitting the photograph did not *actually* harm her friend. I was not rooting around on the off-chance that maybe my wallet really *was* in my bag after all.[13]

Perhaps, then, the belief is temporarily forgotten? When I reach for my wallet, perhaps it is that I just do not remember that it is not with me. When I hesitate before the fudge, perhaps I have just lost track of the fact that it is not dog feces. When I step timidly on the walkway, perhaps I have just forgotten that it is solid. Perhaps. But I do not think this could be the full story. Rozin's subjects hesitate to eat the soup even if they are vividly and occurrently entertaining the thought 'this is a completely sterile bedpan,' fully, consciously and with explicit attention to its meaning and implications. I was rooting around in my bag for my wallet at the exact moment that I was vividly and occurrently entertaining the thought 'I left my wallet in New Haven,' fully, consciously, with explicit attention to its meaning and implications. And certainly the Hulapai Canyon steppers have not *forgotten* that the platform is safe, else they

would do something a good deal more dramatic than hesitate before taking the next step.

But if it is not a case where the subject is deceiving others, or self-deceived, or uncertain, or forgetful, then why *is* stepping onto the Skywalk different from stepping onto the back porch? The reason, of course, is that each activates a different set of affective, cognitive, and behavioral association-patterns. When the subject steps onto the wooden porch, input to her visual system affirms her explicit conscious belief that the surface is solid and secure; this sets into motion a train of associations and activates a number of motor routines. But since these motor routines coincide with those activated by her explicit intention to walk across a surface that she believes to be solid, there is no belief-behavior mismatch. When she steps onto the glass platform, by contrast, input to her visual system suggests that she is striding off the edge of a cliff. This visual input activates a set of affective response patterns (feelings of anxiety) and motor routines (muscle contractions associated with hesitation and retreat), and the visual-vestibular mismatch produces feelings of dizziness and discomfort, leading to additional activation of motor routines associated with hesitation and withdrawal.[14] These motor routines compete with those activated by her explicit intention to walk across a surface that she believes to be solid; the result is the belief-behavior mismatch adverted to above.

Nor do we need anything so dramatic to make the point. The same phenomenon occurs when I set my watch five minutes fast. The effectiveness of the strategy does not depend on my *forgetting* that the watch is inaccurate, or on my *doubting* that it is really 9:40 rather than 9:45, or my *deceiving* myself or others into thinking that it is five minutes later than it is. Rather, as with the glass-bottomed Skywalk, when I look at my watch, input to my visual system suggests that I am in a world where the time is *t*+5. This visual input activates a set of affective response patterns (feelings of urgency) and motor routines (tensing of the muscles, an overcoming of certain sorts of inertia), leading to the activation of behavior patterns that would not be triggered by my explicit, conscious, vivid, occurrent belief that it is actually only 9:40.[15] The activation of these response patterns constitutes the rendering occurrent of what I hereby dub a *belief-discordant alief*. The alief has representational-affective-behavioral content that includes, in the case of the Skywalk, the visual appearance as of a cliff, the feeling of fear and the motor routine of retreat.[16] Similar appeal to belief-discordant alief can be made in each of the other cases. The visual appearance of the feces-shaped fudge renders occurrent a belief-discordant alief with the content: 'dog-feces, disgusting, refuse-to-eat'—an alief that runs counter to the subject's explicit belief that the object before her is composed of a substance that she considers delicious and appealing. The visual-motor input associated with throwing a dart at a representation of a loved one renders occurrent a belief-discordant alief with the content: 'harmful action directed at beloved, dangerous and ill-advised, don't-throw'—an alief that runs counter to the subject's explicit belief that damaging a representation has no effects on the entity represented. The visual-motor input associated with handling cash rendered occurrent my belief-discordant alief with the content: 'Bunch of money. Needs to go into a safe place. Activate wallet-retrieval motor routine now'—an alief that ran counter to my explicit belief that my wallet was in Connecticut while I was in Maryland. And so on.

I.2. A Provisional Characterization of Alief. In the remainder of the article, I argue for the importance of recognizing the existence of alief—so-called because *a*lief is *a*ssociative, *a*utomatic, and *a*rational. As a class, aliefs are states that we share with nonhuman *a*nimals; they are developmentally and conceptually *a*ntecedent to other cognitive attitudes that the creature may go on to develop. And they are typically also *a*ffect-laden and *a*ction generating.[17] I will argue that any theory that helps itself to notions like belief, desire, and pretense needs to include a notion like alief in order to make proper sense of a wide range of otherwise perplexing phenomena. Without such a notion, I will contend, either such phenomena remain overlooked or misdescribed, or they seem to mandate such a radical reconceptualization of the relation between cognition and behavior that traditional notions like belief seem quaint and inadequate. In short, I will argue that if you want to take seriously how human minds really work, and you want to save belief, then you need to make conceptual room for the notion of alief.

Because alief is a novel notion, introduced to make sense of a cluster of otherwise baffling cases, most of the paper will proceed by examination of specific examples. The heart of the paper lies in that discussion, and in the claim that consideration of such cases brings to light

issues of philosophical importance. At the same time, I will tentatively offer a more abstract characterization of the concept that I am introducing, so that the general claim that I making can be properly assessed.

The account that follows is explicitly provisional. I have little doubt that I have gotten some of the details wrong—and perhaps a good deal more than the details. But it seems to me better to make an honest mistake by attempting to be precise than to avoid error by refusing to be explicit. With that in mind, I offer the following tentative characterization of a paradigmatic alief:

> A paradigmatic *alief* is a mental state with associatively linked content that is representational, affective and behavioral, and that is activated—consciously or nonconsciously—by features of the subject's internal or ambient environment. Aliefs may be either occurrent or dispositional.

Nearly every clause in this characterization merits a quick remark or highlighting:

(1) Alief is a *mental* state. . .

Since I incline towards physicalism, this means that I think alief is also a physical state. But it is a special sort of physical state—one that occurs in the brain of a conscious subject. And it occurs in her brain as the result of her (or her genetic ancestors) having undergone certain sorts of experiences—experiences that result in the creation of clusters of associations with representational-affective-behavioral content.

(2) Alief is a mental *state*. . .

Alief is a state and not, say, an attitude. It is (I think) roughly what Aristotle would call a *hexis*.

(3) . . . with *associatively linked* content . . .

That is, a cluster of contents that tend to be co-activated. The contrast here is with discrete contents that fail to be linked through such an association.

(4) . . . that is *representational, affective, and behavioral* . . .

In paradigmatic cases, an activated alief has three sorts of components: (a) the representation of some object or concept or situation or circumstance, perhaps propositionally, perhaps nonpropositionally, perhaps conceptually, perhaps nonconceptually; (b) the experience of some affective or emotional state;[18] (c) the readying of some motor routine.[19]

(5) *Paradigmatic* alief is a mental state with content that is representational, affective, and behavioral . . .

Notwithstanding the characterization offered in (4), I do not want to rule out the possibility of there being aliefs that involve the mental activation of a different sort of associative cluster. Perhaps there are cases where the activation occurs at a sufficiently low level to render the notion of representation inapplicable. Perhaps there are states that lack an obvious affective ingredient, or that do not include the clear activation of a motor routine, but that nonetheless sufficiently resemble our paradigm cases that we want to count them as aliefs. Perhaps there are cases where the most noticeable associations are not easily subsumed under the three categories offered—cases that primarily involve the heightening or dampening of certain sorts of attention, or the heightening or dampening of certain perceptual sensitivities.

(6) Alief is a *mental state with . . . behavioral . . . content.*

That is: alief itself does not involve the *execution* of these motor routines; it merely involves their *activation* (alief is a *mental* state). At the same time, this activation renders it more likely that the routine will actually be performed.[20]

(7) Alief . . . content . . . [may be] activated . . . *consciously or nonconsciously.*

That is: a subject may (occurrently) alieve something with or without being aware of being (put into) in such a state.

(8) Alief . . . content . . . [may be] activated . . . via *features of the subject's internal or ambient environment.*

That is: the activation of an alief may be the result either of (conscious or nonconscious) (quasi-)perception, or of (conscious or nonconscious) nonperceptual thought.[21]

(9) Aliefs may be either *occurrent or dispositional.*[22]

A subject has an *occurrent alief* with representational-affective-behavioral content *R-A-B* when a cluster of dispositions to entertain simultaneously *R*-ish thoughts, experience *A*, and engage in *B* are activated—consciously or unconsciously—by some feature of the subject's internal or ambient environment. A subject has a *dispositional alief* with representational-affective-behavioral content *R-A-B*

when there is some (potential) internal or external stimulus such that, were she to encounter it, would cause her to occurrently alieve *R-A-B*.[23]

(10) Tentative characterization . . .

Despite all that I have said in this section, I continue to waver on whether it would be better to think of the term as two-place (*S* alieves *R*) rather than four-place (*S* alieves *R-A-B*) relation. Had I opted for the former, I might have introduced the expression as follows:

> *S* (occurrently) alieves *R* when *S*'s *R*-related associations are activated and thereby rendered cognitively, affectively and behaviorally salient.

In most of the discussion that follows, I will make use of the expression in its four-place version, occasionally noting cases where the two-place version seems more appropriate.

I.3. Examples and Usage. How does the terminology just introduced help us with our opening examples? Consider, for example, Rozin's subject who shows reluctance to put a piece of vomit-shaped rubber in her mouth. When the visual experience as of vomit awakens in the subject the entertainment of vomit-related trains of thought, the affective experience of disgust, and the activation of motor routines associated with behaviors like retreat and avoidance, Rozin's subjects come to *alieve occurrently* the representational-affective-behavioral content: 'Vomit! Disgusting! Stay away!'[24] And anyone whose inclinations to feel disgust and avoidance would be activated by encountering a vomit-like visual stimulus (a class which for evolutionary reasons is likely to include nearly everyone) *dispositionally alieves* what Rozin's subjects occurrently alieve.

Of course, occurrently *alieving* 'Vomit! Disgusting! Stay away!' is fully compatible with occurrently *believing* that there is no vomit in one's vicinity. An occurrent alief whose content is *P* may well be accompanied by an occurrent belief whose content includes not-*P*. Indeed, it is precisely when they are belief-discordant that aliefs tend to be evident to us. It is because Rozin's hesitating subjects occurrently believe something like: 'the object in front of me is made of sterilized rubber and poses no risk to my health' that we need to explain their reluctance in terms of their alief. (Actually, I think that alief plays a major role in explaining behavior even when it is belief-concordant, an issue to which I return briefly in the closing section. But since the most convincing cases are those involving belief-discordant alief, I will focus primarily on those in making my initial argument.)

One final remark concerning usage. Given that I have opted for the four-place characterization, I need to say that Rozin's subjects occurrently alieve something like 'Vomit! Disgusting! Stay away!' while believing that there is no vomit in their vicinity. Had I opted for the two-place characterization, I might have said instead: Rozin's subjects believe that that the object before them is a piece of rubber, but they alieve that it is a mound of vomit. This usage seems particularly tempting in cases where the associational clusters are awakened by the presence of a particular object or situation, and where the associations awakened tend to be similar across individuals. Indeed, there is a natural tendency to loosen usage yet further, saying, for example, that visitors to the Skywalk believe that the glass surface is safe, but alieve that it is dangerous; that Rozin's dart-throwers believe that damaging the picture will not harm their loved one, but alieve that it will; that Rozin's shirt-avoiders believe that their enemy's laundered chemise is utterly harmless, but alieve that wearing it is ill advised; that Charles believes that he is at no risk from the slime, but alieves that it is about to attack him. I consider it a live possibility that careful reflection on natural patterns of usage will reveal that I have made the wrong decision in opting for the four-place characterization. But for the time being, I will explore the advantages of employing the term in the way that I have characterized it thus far.

This ends the official introduction of the notion of alief. In the remainder of the paper, I do three things. In section II, I offer some brief additional general remarks about the relation between the state of alief and propositional attitudes such as belief, desire, and pretense. In section iii, I offer a series of examples—drawn from recent empirical work in psychology—that played a central role in convincing me that appeal to the notion of alief is crucial if we wish to hold on to a notion like belief that relates to action in anything like the way philosophers have traditionally assumed. In section iv, I close with a few speculative remarks about ways that appeal to the notion of alief may be help us to make sense of two apparently unrelated phenomena: the tendency of examples to affect us in ways that abstract descriptions do not; and the role of habit in Aristotelian ethics.

II. Alief and Other Attitudes

II.1. Alief, Belief and Imagination. Why can't alief be assimilated to one of the more familiar cognitive attitudes—belief, for example, or imagining? There are a number of reasons that I think that it cannot, which I will present in the remainder of this section.[25]

Alief differs from both imagining and believing along certain crucial dimensions. If I believe that *P*, I believe that it is true that *P*, and my belief is nondefective only if, as a matter of fact, it is true that *P*. If I suppose or imagine or pretend that *P*, I suppose or imagine or pretend that it is true that *P*, but the actual truth or falsity of *P* is explicitly irrelevant to my successfully supposing or imagining or pretending it to be. Both classes of states, then, involve what Velleman helpfully calls *accepting* a proposition: to believe or imagine or suppose or pretend that *P* is to regard *P* as true (in some way[26]). But though they coincide in this dimension, they differ in another: whereas belief is *reality-sensitive*, supposition and imagination and pretense are explicitly *reality-insensitive*. It is this latter disparity that is typically taken to underlie one important difference between belief on the one hand, and supposition, imagination, and pretense on the other: whereas (modulo certain complications) we can imagine pretty much any content, we can (without acrobatics) believe only what we take to be true.

How does alief fare along these dimensions? Strictly speaking, it lies in another plane altogether. Believing and supposing and imagining and pretending are all (at least on certain uses of the expressions in question) propositional attitudes, whereas alieving (as I am provisionally using the expression) is not. But we can, by employing the 'loose' usage adverted to above, make reasonable sense of the notion of alieving that *P*, and we can ask—keeping in mind that our usage is loose—whether alieving that *P* involves accepting that *P*. We will need to be a bit more careful when we ask whether alief is reality-sensitive or reality-insensitive, and whether we are in a position to alieve at will. But again, we will be able to draw certain fairly sharp contrasts between alief and other attitudes.

Let us begin with the question of acceptance. Does alieving that *P* involve accepting that *P*? (That is, does being an alief state with the content *R-A-B* involve regarding it as true in some way that *R* is part of one's real or imagined environment?[27]) Interestingly, the answer to this question turns out to be: *no*, and the way in which it turns out to be *no* reveals something important about the nature of alief. Unlike belief or pretense or imagination or supposition, alief does not involve acceptance. Though the point can be made on conceptual grounds alone, it is helpful to begin with a specific example.

In a 1986 study by Rozin, subjects saw 'sugar poured into two bottles, and then applied labels of *sugar* and *sodium cyanide*, each to one of the bottles, making their own choice.' Despite having applied the labels themselves, subjects 'showed a reluctance to consume sugar from the cyanide-labeled bottle.'[28] So far, the case is a familiar one: while Rozin's subjects believed that both bottles contained sugar, consideration of the second rendered occurrent an alief state with the content 'cyanide, dangerous, avoid' associated with the second bottle—and this belief-discordant alief played a role in governing their behavior.[29] Up to this point, there is no reason to posit a case of alief without acceptance: in alieving 'cyanide, dangerous, avoid' the subject is regarding as true (perhaps in imagination) that the bottle contains cyanide.

The interesting case comes from a follow-up study four years later. In that study:

> Subjects faced two empty brown 500 ml bottles. In the presence of the subject, the experimenter opened a container of Domino cane sugar, and poured some into each bottle, so that about ¼ of each bottle was filled. The experimenter informed subjects that she was pouring sugar into each bottle. The experimenter then presented the subject with two typed labels. One had *not sodium cyanide, not poison* written on it, with a red skull and cross bones preceded by the word *not*. The other label had *sucrose, table sugar* typed on it. The subject was invited to put one label on each bottle, in any way he or she chose. The experimenter then set out two different colored plastic cups, one in front of each bottle, and poured unsweetened red (tropical punch) Kool-Aid from a glass pitcher into both, until they were about half full. Now, using separate, new plastic spoons for each bottle, the experimenter put a half spoonful of powder from one sugar bottle into the glass standing in front of that bottle, and repeated this with the other glass for the other sugar bottle (ibid.).

Subjects then faced the choice of drinking from the cup containing the sugar that had been labeled 'sucrose, table sugar' or from the cup containing the sugar that had been labeled 'not

sodium cyanide, not poison.' Though the effect was somewhat less pronounced than in the original study, subjects showed considerable reluctance to drink from the latter.

Here again, while Rozin's subjects believed that both bottles contained sugar, consideration of the second bottle rendered occurrent an alief state with the content 'cyanide, dangerous, avoid.' But in this case, the label read precisely the opposite: it 'had *not sodium cyanide, not poison* written on it, with a red skull and cross bones preceded by the word *not.*' So, although these subjects were in an alief state with the content 'cyanide, dangerous, avoid,' the content they were prompted to imagine was exactly the opposite. They did *not*—as the acceptance condition requires—regard it as true in some way that cyanide is to be found in the vicinity; instead, it was the negated presence of the word 'cyanide' that rendered occurrent their cyanide-associated aliefs.

Can we explain this with the resources of only belief and imagining? Clearly, belief cannot do the work: it is implausible to suggest that the subject believed that the bottle she had labeled 'not sodium cyanide, not poison' contained cyanide. But what about imagining? Can't we say that the source of the subject's hesitation is that she first imagines that the bottle does contain poison, and that she then somehow negates this, and that this enables her (perhaps in some special Sartrean fashion) to imagine the absence of poison?[30]

Perhaps this is indeed what happens. But how is this supposed to explain the subject's hesitancy to drink the liquid? Is the reason for her hesitancy supposed to be that she *had been imagining* that the bottle contained cyanide, though now she is not—and that what she imagined *in the past* (though fails to imagine now) somehow explains her action *at present*? Or that her current imagining that the bottle does not contain cyanide somehow contains within it (in not-fully-*aufgehoben* form) the antithetical imagining that the bottle does contain cyanide? And that somehow this negated semi-imagined content—content that she has, throughout the entire process, been fully consciously aware of explicitly disbelieving—sneaks into the control center for her motor routines and causes her to hesitate in front of the Kool-Aid?

Really? Is this really what you think imagining is like? Or have you just described a case of belief-discordant (and imagination-discordant) alief: a case where the subject believes that the bottle does not contain cyanide, imagines that the bottle does not contain cyanide, yet has an occurrent alief with the content: cyanide, dangerous, avoid? Is it not a lot more natural to describe this as a case of alief-motivated behavior than as a case of motivation by (past or negated) imagination? And if it is alief that is doing the explanatory work here, is it not plausible that alief is doing the explanatory work in the cases above as well?

For those unconvinced by examples or lines of rhetorical questioning, there is a more general argument for why alief can occur without acceptance. At its core, alief involves the activation of an associative chain—and this is something that can happen regardless of the attitude that one bears to the content activating the associations. (Indeed, since alief may be activated nonconsciously, one may bear towards that content no attitude at all.) This means that alief contexts are what we might call *hyperopaque*: they do not permit *salva veritate* substitution even of expressions that the subject explicitly recognizes to be coreferential.[31] Even if I believe that the phrases 'not poison' and 'safe to consume' pick out coextensive classes of substances, even if I focus on that belief and hold it vividly before my mind, even if the synonymy of these two terms is crucial to my views about some other matter, still the aliefs activated by the two expressions may be wildly dissimilar.[32] Imagination, by contrast, is not hyperopaque in this way. If I explicitly recognize that P and Q are synonymous, and I imagine P while focusing explicitly on the co-referentiality of P and Q, then in imagining P I imagine Q. Alief just is not imagination.

The same features that explain alief's hyperopacity and the possibility of alief without acceptance explain why we are not in a position to alieve at will. If I believe that P, and subsequently learn that not-P, I will revise my belief. If I imagine that P, and subsequently learn that not-P, I will make no such revision. But what if I (loosely speaking) alieve that P, and subsequently learn that not-P? What happens then? At first glance, alief seems to behave like imagination and its kin: after all, the cases above are all cases where the subject truly and consciously believes P while actively alieving not-P. But this does not quite capture the full story. If I believe that P and imagine that not-P, I am violating no norms. But if I believe that P and alieve that not-P, something is amiss. Learning that not-P may well not cause me to cease alieving that P—but if it does not, then (though other considerations may override this) I am violating certain norms of cognitive-behavioral

coherence. No such criticism is possible in the analogous case of imagining.

To the extent that action is supposed to be responsive to reality, the well-functioning aliever is one whose aliefs and beliefs largely coincide (or one whose ability to suppress contrary impulse is strong[33]). But alief just is not reality-sensitive in the way belief is. Its content does not track (one's considered impression of) the world. At the same time, it is not reality-insensitive in the way that imagination is. For while we can (for the most part) imagine at will, we do not seem to have the same sort of freedom in alief.[34] We may be relatively unconstrained in which of our dispositional aliefs we render occurrent—at least in the case of those aliefs that can be rendered occurrent through contemplation alone—but we are far from unconstrained in which dispositional aliefs we have in the first place. Our dispositional aliefs depend on the associational patterns that have been laid down in our minds as the result of our experiences and those of our genetic ancestors. We are not in a position to generate such patterns of association merely at will.

So it looks like, just as it is (something close to) *conceptually* impossible to believe at will, it is *practically* impossible to alieve at will. Of course, in both cases we might use all sorts of tricks to bring ourselves to be in a certain sort of mental state—'roundabout routes' involving processes that we ourselves deliberately initiate.[35] But if we use such tricks to cultivate beliefs, we need to cover our tracks;[36] if we use them to cultivate aliefs, we can do so under conditions of full disclosure.

This concludes the brief survey contrasting alief with attitudes like belief and imagining. We now turn to the second issue of this section, the relation between these attitudes, and the bringing about of behavior. I will suggest that alief's special structure—its being a mental state with affective, representational, and behavioral content that is activated by features of the environment—means that it poses problems for behavioral accounts of belief that are especially severe.

II.2. Alief and Behavior. According to what Velleman has dubbed the 'purely motivational view of belief,' 'all that's necessary for an attitude to qualify as a belief is that it disposes the subject to behave in certain ways that would promote the satisfaction of his desires if its content were true. An attitude's tendency to cause behavioral output is thus conceived as sufficient to make it a belief.'[37] Or, again: to believe that *P* is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which *P* (together with one's other beliefs) were true.[38]

There are at least three sorts of marginal cases where this sort of analysis seems to go awry—two that pose problems for necessity, the third for sufficiency. The first sort are cases where (arguably) a subject believes that *P*, but where this belief does not bring with it a disposition to act in *P*-concordant ways because of some feature of the subject. (Think, for example, of an immutable omniscient purely contemplative God, a permanent paralytic, a subject built to act with utter randomness, a character under an unbreakable spell that causes him to act contrary to his first-order intentions, a hopeless akratic, or an agent who aims always to deceive.) The second are cases where (arguably) a subject believes that *P*, but where this belief does not bring with it a disposition to act in *P*-concordant ways because the belief itself has no behavioral implications. (Think, for example, of a subject who believes in causally-inert invisible goblins, or of a subject who believes that she inhabits a space that is distorted but Euclidian (rather than undistorted but non-Euclidian).[39]) The third are cases where, although the subject is disposed to act in the requisite ways, she nonetheless fails to believe that *P* because she lacks beliefs (either locally or globally). (Think, for example, of a super-stoic who acts and has desires but always withholds assent, or of a hyper-Van Fraassenite who extends his constructivist commitments to the realm of the observable.)

Five-finger exercises that they are, these marginal cases do not show that there is anything deeply wrong about the motivational view. All that is needed to avoid them are a few tweaks to the notion of disposition and a reiteration of the irrelevance of mental states. The big guns come loaded with a different sort of ammunition: not with the suggestion that the view is wrong in certain far-fetched contrived cases, but with the assertion that it is problematic through and through because of a wide range of attitudes—among them acceptance (Michael Bratman), imagination (Gregory Currie, Velleman) and pretense (Tyler Doggett and Andy Egan, Velleman)—may motivate *P*-concordant behavior.[40]

Even here, I think there is room for the defender of a neo-behaviorist account. Restrict yourself to nondeviant subjects, and retreat, say, to betting behavior or high-stakes situations. Once again, you can save the letter of the view that belief and behavior go hand in hand.

To some extent, this strategy works for alief as well. (If it did not, it would be hard to maintain that the paradigmatic cases above are ones in which the subject believes that *P* but alieves that not-*P*.) H. H. Price, whose underappreciated discussion of related examples deserves more detailed attention than I have space for here, employs such a strategy. Defending his account of a case of what he calls 'half-belief,' Price writes:

> It might be suggested that the man who avoids walking under ladders does just *believe* (however unreasonably) that walking under ladders has bad consequences . . . . After all, these people act as if they believed, and they often go to considerable trouble in consequence. They step off the pavement into a muddy street or even into a street full of traffic, to avoid the ladder . . . . Moreover they show the emotional symptoms of belief, for example, discomfort or unrest if there is . . . no way of avoiding the ladder. . . . Of course, these people will not admit that they . . . believe these propositions; not even to themselves, and still less in public. . . . But one may hold beliefs . . . without admitting to oneself that one holds them (op. cit., p. 310).

Price rejects this account—a proposal, he suggests, to 'dispense with the concept of half-belief altogether'—because, while

> . . . no doubt there are some who do wholly believe that their chances of suffering misfortunes are increased if they walk under a ladder . . .. I do not think that this is the usual situation, . . . the ordinary person who avoids walking under ladders does not seriously believe that walking under ladders does any harm, or at any rate he does not believe it with complete seriousness. We notice that if it is very important for him to get to his destination quickly (for example, if he will miss a train if he does not hurry) he does not seem to mind the ladder at all. He sees it—there it is, in front of his nose—but he goes straight under it without hesitation. He himself, if he thinks about his experience afterwards, will be able to notice that he felt no qualms at all about doing the thing which he ordinarily avoids so carefully (op. cit., pp. 310–11).

'A half-belief,' he concludes, is 'something which is 'thrown-off' when circumstances alter. . .. [I] n some contexts to which the proposition is relevant one is in a belief-like state about it, but in other contexts to which it is equally relevant one disbelieves it or disregards it.' This is so even though 'in both sorts of contexts, the evidence for the proposition . . . remains the same, and the probability of the proposition is as great, or as little, as it was before' (op. cit., p. 312).

I agree with Price that the ladder case might well proceed as he describes. But I am not so clear that his analysis will work for the cases presented on the opening pages. Suppose it is very important for me to get to my train, but that the station lies across a chasm fifty feet wide and 1,000 feet deep, bridged by a transparent glass walkway. Even if I 'will miss a train if [I do] not hurry,' I do not think it is true that I would 'not seem to mind the [apparent chasm] at all,' crossing it 'without hesitation' even though the visual stimulus is 'right under my nose.' I very much doubt that in 'think[ing] about [my] experience afterwards,' I would 'be able to notice that [I] felt no qualms at all about doing the thing which [I] ordinarily avoid so carefully.' (Indeed, in my own case, I am not sure I could make it across the bridge at all without closing my eyes—which would be, of course, to suspend the occurrent alief by suspending the feature that activates it.)

Suppose we raise the stakes. My child is on the other side of the chasm, and I need desperately to reach him to prevent some dreadful occurrence. Here I suspect I could make it across the bridge—eyes open—to perform the rescue: after all, I believe that he is in danger, and I believe that the bridge is safe. But even here, the hesitation would not fully dissipate. And not because I doubt in any way that the surface is sturdy: I see others walking across it and am about to do so myself. I am 100% certain that I will make it safely—as certain as I would be if the chasm were only 5 feet deep, as certain as I would be if the bridge were made of opaque material. Still, I hesitate; still, I shudder. My behavior reflects something other than my belief. It is my alief in action.

The reason Price's explanation fails for our paradigm cases is that the mechanisms they exploit are not under our direct control. We are not in a position to 'throw them off . . . when circumstances alter.' This is not because we are in doubt about what we believe. There is no question in my mind that the fudge has not been transformed into dog feces; there are few things of which I am more certain than that hurling darts at a photo of my baby will do no harm to the baby itself. Still—even in high-stakes situations—there is a hesitation to my belief-concordant actions.[41]

The problem with the belief-behavior picture is that at its heart lies a faulty picture of what makes us act.[42] I do not doubt that the account

could be made extensionally adequate: limit the cases that count as 'behavior' in the relevant sense, fuss with the notion of disposition, make the fate of the world depend on the subject's actions. Belief and behavior can be made to match up, so long as one is free to make relevant alterations from both directions. But deep down, the account misses something very important about human behavior. This is something to which both Aristotle and Hume were especially well attuned (I will return to this in the final section), and which contemporary psychology has begun to explore in detail. It is to cases from the latter domain that I turn in the next section.

III. Automaticity

Recent work on 'automaticity' has produced a remarkable series of widely publicized results suggesting that alief plays a larger role in behavior than many had thought. Indeed, one of the main projects in social psychology over the last two decades has been to document systematically the ways that behavior-inducing mental representations may be activated by awakening the associative patterns that have come to be linked with some object, stereotype, protocol, or mental image.[43] A few examples will suffice for giving a sense of their flavor. But it is important for the reader to realize that this is a massive research program and that while it may be possible to come up with alternative explanations for one or another of the examples I discuss, the basic phenomenon I am describing here has been established beyond any reasonable doubt in hundreds of published studies.[44]

Much of the work in this area has been pioneered by John Bargh and his colleagues, who, in a typical task present subjects with some sort of association-inducing stimulus. This is often a 'scrambled sentence' task—a standard technique in psychology used to 'prime' particular concepts.[45] In one such study, subjects faced one of three conditions: either the collections of words from which they were asked to form sentences contained only neutral terms, or they also contained a number of terms associated either with politeness (for example, respect, honor, considerate, patiently, courteous) or rudeness (for example, aggressively, bother, disturb, intrude, brazen). Subjects were instructed that, after completing the task, they should come out into the hallway and find the experimenter, who would then give them the next task to complete. When they emerged, they found the experimenter engaged in a conversation with another 'subject' (actually a confederate), a conversation that continued either until the first subject interrupted the conversation, or until 10 minutes had passed.

The action-patterns of the three groups differed markedly. Of those who had been primed with the rudeness concept, most interrupted in the allotted time; those in the neutral condition interrupted in less than half of the cases; whereas those in the polite condition interrupted in almost none of the cases.[46]

One *might* maintain that the various groups differ in their beliefs, or that they differ in their desires, or that the subject's interruption of the experimenter is not an action of the sort that belief-desire explanations are designed to cover. I have no doubt that such a story could be told. One might say, for instance, that all three groups share the same desire—to interrupt the experimenter only if doing so would be socially acceptable—but that they differ in their belief about whether it is. (Note that this would involve attributing to the subjects an odd sort of belief—one that is formed as the result of mechanisms that are not themselves sensitive to any subject-independent truth attitudes.[47]) Alternatively, one might try to explain the phenomenon in terms of imagination or pretense. Perhaps engaging in the scrambled sentence task causes the subjects to fantasize that the experimenter is rude, or polite—or that they themselves are rude or polite—and, carried away by this fantasy, perhaps they begin to act as if it were true. Perhaps.[48] But why would engaging in the scrambled sentence task cause the subjects to engage in this sort of fantasy (unless, of course, the explanation runs through something like the notion of alief)? And even if we have an answer to that question, why would engaging in such a fantasy make them act as if it were true (again, unless the explanation runs through something like alief)?

Rather, what Bargh and his colleagues have done, I want to argue, is to induce in their different sets of subjects different sorts of occurrent alief. As the result of the pre- or quasi-conscious activation of the cluster of affective tendencies and behavioral repertoires associated with the notion of rudeness, subjects in the third condition find themselves more likely to act in ways that they would act in the presence of rudeness; as the result of the pre- or quasi-conscious activation of the cluster of affective tendencies and behavioral repertoires associated with the

notion of politeness, subjects in the second condition find themselves more likely to act in ways that they would act in the presence of politeness.[49]

Nor is this an isolated anomaly. Example after example reveals the subtle role of alief in guiding behavior. In another widely publicized Bargh experiment, subjects performed a scrambled sentence task in which one group confronted sentences containing terms associated with the elderly (for example, wrinkle, bingo, and retired), whereas the second group's unscrambling task involved only neutral terms. After completing the experiment, subjects were surreptitiously timed as they walked down the hall to the elevator. Those primed with the elderly stereotype took significantly longer to walk to the elevator than those who had not been so primed.[50]

It seems implausible (to say the least) that Bargh's elderly-primed subjects *believed* that they had suddenly turned into a bunch of geezers who needed to dawdle lest they overtax themselves. It is slightly less absurd to suggest that Bargh's elderly-primed subjects *imagined* themselves as old—or imagined someone else who is old—and, having so imagined, began to act in some ways as if the imagined content should govern their own actual behavior. But even this is a rather far-fetched explanation.[51] (Among other things, in well-designed scrambled sentence tasks, subjects remain unconscious of the fact that a particular notion is being primed.[52]) Rather, I want to suggest, Bargh's elderly-primed subjects occurently *alieved* below the level of conscious awareness something like: 'Old. Tired. Be careful walking to that elevator . . .'—and the activation of this behavioral repertoire made them more likely to act in accord with it.

Additional research within this paradigm has reinforced and expanded the lessons of these early experiments. So, for example, showing suitably primed subjects a picture of a library leads them to speak in quieter tones; showing them an image of an elegant dining room—or exposing them to the smell of soap—leads them to eat more neatly.[53] Subliminal visual priming with an image of an African-featured face leads subjects to respond more aggressively to certain sorts of provocation.[54] Priming subjects with thoughts of their (achievement-oriented) mother leads them to persist longer at word-find tasks; priming them with thoughts of a friend makes them more likely to help a stranger.[55]

Indeed, alief may be activated in even more striking ways. Recently, psychologist Lawrence Williams 'hypothesized that a simple experience of physical, spatial distance would trigger feelings of psychological distance and that those feelings, in turn, allow people to enjoy aversive media.' Subjects were first asked to plot a pair of points on a cartesian plane: the points were either quite close to one another (occupying less than 1/4 of the plane) or quite far apart.

> All the participants then read an embarrassing passage from a novel—in which a woman opens a magazine to find that her ex-boyfriend has written an article about her, called 'Loving a Larger Woman'—and rated how much they enjoyed the story. Just as Williams had expected, the participants who drew the dots far apart liked the passage more.
>
> In his next study, after the volunteers drew the dots, they read a book excerpt in which a man beats his brother with a rock after a car crash. When the readers rated their emotional experience, Williams found, people who were told to draw the dots close together reported feeling more negative emotions.[56]

In all of these cases, it is perhaps possible to explain what is going on in familiar terminology. Perhaps Bargh's interruption subjects imagine that there is rudeness afoot in their dominion, and adjust their behavior accordingly. (Really? Even though the priming takes place at the unconscious level?) Perhaps his elevator subjects imagine that they are old and gray and full of sleep, and consequently slow their pace. Perhaps Williams's subjects imagine that they are far away from the stories they hear, and therefore feel their emotional tug less strongly.

Perhaps. Or perhaps what is happening in each of these cases is the activation of a low-level cluster of associations—representational, affective, behavioral—an activation that renders the subject more likely to exhibit behavior of a certain sort. To a reasonable approximation, it looks like all depictive representations—even those that we explicitly disavow as false—feed into our behavioral repertoires, and that it is only through a process of conscious or habit-governed inhibition that representations whose accuracy we endorse come to play a distinctive role in governing our actions.

If so, there is something deeply wrong about the traditional picture of the relation between belief and behavior that we discussed in section ii. But of course, this is not the only way philosophers have thought about these matters. In the final section, I briefly examine one competing philosophical strand.

IV. Alief, Persuasion, and Habit

Despite certain protestations to the contrary, philosophers have been exquisitely sensitive to the ways in which contemplation of an imaginary particular may have cognitive and motivational effects that differ from those evoked by an abstract description of an otherwise similar state of affairs.[57] (Think of Plato's cave, the ring of Gyges, twin earth, the Chinese room, teletransportation, Thomson's violinist, the veil of ignorance, Mr. Truetemp, the fat man on the bridge, and any of the myriad other examples). A particularly vivid presentation of this claim can be found in Hume's *Treatise on Human Nature*, where Hume writes:

> There is a noted passage in the history of Greece, which may serve for our present purpose. Themistocles told the Athenians, that he had form'd a design, which wou'd be highly useful to the public, but which 'twas impossible for him to communicate to them without ruining the execution, since its success depended entirely on the secrecy with which it shou'd be conducted. The Athenians, instead of granting him full power to act as he thought fitting, order'd him to communicate his design to Aristides, in whose prudence they had an entire confidence, and whose opinion they were resolv'd blindly to submit to. The design of Themistocles was secretly to set fire to the fleet of all the Grecian commonwealths, which was assembled in a neighbouring port, and which being once destroy'd wou'd give the Athenians the empire of the sea without any rival. Aristides return'd to the assembly, and told them, that nothing cou'd be more advantageous than the design of Themistocles but at the same time that nothing cou'd be more unjust: Upon which the people unanimously rejected the project.[58]

Hume goes on to note that his contemporary Charles Rollin found it astounding that the Athenians would reject—merely on grounds of injustice—a strategy so 'advantageous' that it would give them 'the empire of the sea without any rival.' But Hume himself is not surprised:

> For my part I see nothing so extraordinary in this proceeding of the Athenians. . . . [T]ho' in the present case the advantage was immediate to the Athenians, yet as it was known only under the general notion of advantage, without being conceiv'd by any particular idea, it must have had a less considerable influence on their imaginations, and have been a less violent temptation, than if they had been acquainted with all its circumstances: Otherwise `tis difficult to conceive, that a whole people, unjust and violent as men commonly are, shou'd so unanimously have adher'd to justice, and rejected any considerable advantage (ibid., II.iii.6.4).

Hume's story brings out the way in which engagement of the cognitive mechanisms associated with vivid imagining may lead a subject to reverse a prior commitment, selecting as preferable the option previously rejected, and shunning the option previously embraced.[59]

For the reader who has gotten this far, it should be apparent what lesson I want to draw from this case. Ever sensitive to the role of habit and association—'If any thing can intitle the author to so glorious a name as that of an inventor, 'tis the use he makes of the principle of the association of ideas'[60] —Hume is here pointing out that judgment about a particular case may be driven as much by alief as by belief. Like his K Street counterpart, Hume recognizes the citizen who believes that wealth should be redistributed across generations alieves that the death tax is unfair; like his Madison Avenue foil, Hume recognizes that a customer who believes that a $9.99 scarf costs nearly ten dollars alieves that it costs only nine. When the citizen votes against the amendment does this show that he really opposes redistribution? Or does it show that action is often governed by alief?

If so, then Aristotle is right: In order to live well, we must work to bring our habits in accord with our reflective beliefs:[61]

> Men become builders by building and lyre-players by playing the lyre; so too we become just by doing just acts, temperate by doing temperate acts, brave by doing brave acts . . . states of character arise out of like activities . . . . It makes no small difference, then, whether we form habits of one kind or of another from our very youth; it makes a very great difference, or rather all the difference.[62]

My conclusion should not be a surprising one. I think that alief governs all sorts of belief-discordant behavior—the cases with which I began the paper, and the ones that I have presented along the way. But if alief drives behavior in belief-discordant cases, it is likely that it drives behavior in belief-concordant cases as well. Belief plays an important role in the ultimate regulation of behavior. But it plays a far smaller role in moment-by-moment management than philosophical tradition has tended to stress.

NOTES

* I am grateful to the Yale University faculty lunch group for comments on a very early draft of this paper, and to audiences at Princeton University (March 2007), the Central American Philosophical Association meeting in Chicago (April 2007), and the *Mind & Language* Pretense Conference at University College, London (June 2007) for excellent questions, comments, objections, and suggestions regarding the talk which served as its immediate predecessor. For more recent discussion and comments, I thank John Bargh, Paul Bloom, Richard Brooks, Carolyn Caine, David Chalmers, Greg Currie, Paul Davies, Andy Egan, Roald Nashi, Elliot Paul, Eric Schwitzgebel, Ted Sider, Jason Stanley, Zoltán Gendler Szabó, and Jonathan Weinberg.

I discuss additional aspects of the notion of alief in a companion article, "Alief in Action (and Reaction)," *Mind & Language* XXIII, no. 5 (November 2008): pp. 552–85.

1. Rothstein, "Skywalk Review: Great Space, Glass Floor-Through, Canyon Views," *The New York Times* (May 19, 2007).
2. Indeed, the story is a slight variation on the early modern 'problem of the precipice,' discussed—among others—by Hume (*Treatise* 1.3.13.10, 148), Pascal (*Pensées*, section 44) and Montaigne (*Essays*, Donald Frame, trans., [Redwood City, CA: Stanford University Press, 1957] p. 250). See Saul Traiger, "Reason Unhinged: Passion and the Precipice from Montaigne to Hume," in *Persons and Passions: Essays in Honor of Annette Baier*, Joyce Jenkins, Jennifer Whiting, and Chris Williams, eds., (Notre Dame, IND: Notre Dame: University Press, 2005), pp. 100–15. I discuss precipice cases in more detail in Gendler (op. cit.).
3. The *physiological* explanation, of course, is that there is a mismatch in input between the visual, vestibular and somatosensory systems. For discussion, see Thomas Brandt and R. B. Daroff, "The Multisensory Physiological and Pathological Vertigo Syndromes," *Annals of Neurology* VII, no. 3 (1980): pp. 195–203; and Thomas Brandt, *Vertigo: Its Multisensory Syndromes* (New York: Springer, 1999/2003), 2nd ed.
4. Throughout my discussion, I am using the term 'content' in a somewhat idiosyncratic way, for want of a better term to describe the general notion that I wish to capture. As I am using the term, content need not be propositional, and may include—as the example above makes clear—affective states and behavioral dispositions.
5. Cf. J. G. Frazer. *The New Golden Bough: A Study in Magic and Religion* (abridged) (New York: Macmillan, 1959; edited by T. H. Gaster, 1922; original work published 1890); Marcel Mauss. *A General Theory of Magic*, Robert Brain, trans., (New York: Norton, 1972; original work published 1902) (as cited in Paul Rozin, Linda Millman, and Carol Nemeroff, "Operation of the Laws of Systematic Magic in Disgust and Other Domains," *Journal of Personality and Social Psychology* L, no. 4 [1986]: pp. 703–12).
6. Rozin, Millman, and Nemeroff, op. cit.
7. The descriptions of the cases make it clear that the experimenters go out of their way to avoid the possibility of any sort of confusion. In the fudge study, for example, 'subjects were offered a piece of high-quality chocolate fudge, in a square shape, on a paper plate [and then] ate the piece . . . . [Next] two additional pieces of the same fudge were presented, each on its own paper plate.' Subjects were made explicitly aware that the two pieces come from the same initial source, and that the only difference between them is that 'one piece was shaped in the form of a disc or muffin, the other in the shape of a surprisingly realistic piece of dog feces.' Despite recognizing that they contained identical ingredients, subjects showed a striking reluctance to consume the feces-shaped piece. See Rozin, Millman, and Nemeroff, op. cit., p. 705.
8. For definition, see: http://en.wikipedia.org/wiki/Cooties. Apparently, a roughly equivalent British term is 'lurgi.'
9. Legally, one is *not* required to carry identification in order to fly. Rather, the Transportation Safety Administration (TSA) requires that airline passengers either 'present identification to airline personnel before boarding or be subjected to a search that is more exacting than the routine search that passengers who present identification encounter.' Cf. *Gilmore v. Gonzales*, 04-15736 D.C No. CV-02-03444-SI Opinion. (Full text at http://www.ca9.uscourts.gov/ca9/newopinions.nsf/A6AE4C85241C517C88257101007B72EB/$file/0415736.pdf ?openelement.) As a quick Internet search for 'flying without identification' will reveal, however, there is a gap between the law and the practice: there were, no doubt, additional features of my particular circumstance that led me to be offered this option.
10. Kendall Walton, "Fearing Fictions," *The Journal of Philosophy*, LXXV, no. 8 (January 1978): pp. 5–27, see p. 5.
11. Although belief is clearly one of the central notions in epistemology, the question of what belief *is* has been (with important exceptions) underexplored in this context. (Of course, there have been extensive discussions of this question in the context of philosophy of mind (for an overview, see section 1 of Eric Schwitzgebel, "Belief", *The Stanford Encyclopedia of Philosophy* (Fall 2019 edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2019/entries/belief/ (first published in the Fall 2006 edition). But (with some important exceptions) this literature has remained largely insulated from the literature in epistemology). One might think a simple characterization would suffice—something like: 'To believe a proposition is to hold it to be true' (Simon Blackburn, *The Oxford Dictionary of Philosophy* [New York: Oxford University Press, 1996], p. 40). But, for reasons that David Velleman brings out nicely (Velleman, "On the Aim of Belief," in *The Possibility of Practical Reason* [New York: Oxford University Press, 2000], pp. 244–82), this will not quite do (at least, not without a careful spelling out of what 'hold to be true' amounts to, which just pushes the question one step back). Moreover, the issue is complicated by there being at least two apparently different fundamental notions of belief: what H. H. Price calls the 'occurrence' or 'traditional' view—that to believe a proposition is to be in a mental state with a particular sort of

introspectively available feature, such as 'vivacity' or 'liveliness' or 'solidity' (a view he attributes to, among others, Descartes, Hume, Spinoza, Cardinal Newman, and Cook Wilson)—and what he calls the 'dispositional' or 'modern' view—that to believe a proposition is to be disposed to act in certain ways (a view he attributes to, among others, Alexander Bain, R. B. Braithwaite, and Gilbert Ryle). See Price, *Belief* (London: Allen and Unwin, 1969). I will have more to say about this matter below. In the meantime—as the astute reader will have suspected by now—I invoke this legacy as much to exculpate as to inform: though I will offer more details in subsequent sections, for the time being, I will leave the notion of belief undefined. (For further discussion, see section 2 of Gendler, "Alief in Action [and Reaction]," op. cit.)

12. I discuss these issues in more detail in Gendler, "Self Deception as Pretense," *Philosophical Perspectives: Mind* (2008).
13. Nor are these cases of what Schwitzgebel ("In Between Believing," *The Philosophical Quarterly* LI, no. 202 [2001]: pp. 76–82) calls 'in-between beliefs'—attitudes 'that are not quite accurately describable as believing that P, nor quite accurately describable as failing to believe that P' (op. cit., p. 76)—cases such as 'gradual forgetting, failure to think things through completely, and variability with context and mood' (op. cit., p. 78). They are closer to some of the cases that Price calls 'half-beliefs' (op. cit., pp. 302–14); I discuss Price's examples in more detail below.
14. For detailed discussion, see Brandt, *Vertigo: Its Multisensory Syndromes*, chapter 29 ("Visual Vertigo: Visual Control of Motion and Balance"), pp. 409–40.
15. Examples of such cases are manifold. I think, for example, that many of the cases of motivation by imagination discussed in David Velleman's "On the Aim of Belief" are actually cases of motivation by alief. Likewise, I think that many of the cases of heuristic-based reasoning discussed by Daniel Kahneman and Amos Tversky are cases of decision on the basis of alief. Cf. Kahneman, P. Slovic, and Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases* (New York: Cambridge University Press, 1982); Kahneman and Tversky, eds., *Choices, Values and Frames* (New York: Cambridge University Press, 2000); cf. also Veronika Denes-Raj and Seymour Epstein, "Conflict between Intuitive and Rational Processing: When People Behave against Their Better Judgment," *Journal of Personality and Social Psychology* LXI, no. 5 (1994): pp. 819–29; and other work in the 'dual processing' tradition. For additional discussion, see Gendler, "Alief in Action (and Reaction)."
16. Of course, stepping onto the wooden deck also renders occurrent an alief—indeed many aliefs—but since those aliefs accord with the subject's explicit beliefs, we do not need to make appeal to them in order to explain her subsequent behavior.
17. An alternative term might be *prelief*, but this expression is already spoken for (cf. J. Perner, S. Baker, and D. Hutton, "Prelief: The Conceptual Origins of Belief and Pretence," in *Children's Early Understanding of Mind*, Charlie Lewis and Peter Mitchell, eds., [Hove, UK: Erlbaum, 1994], pp. 261–86). And in any case, it lacks the resonance of the chosen term. One might also want to leave room for a notion related to desire in something like the way that alief is related to belief. Had 'prelief' been available, one might choose *presire*; since it is not, a suitable expression is *cesire*. (I remain utterly agnostic about what sort of attitude cesire might be.)
18. Our affective processing mechanisms seem to be fairly insensitive to the question of whether the scenario under consideration is real, imagined, supposed or denied. (To the extent that there is a difference in the intensity of our responses, this can be largely traced to a difference in the intensity of the stimulus.) (Cf., for example, the literature surveyed in Anthony R. Damasio, *Descartes' Error: Emotion, Reason and the Human Brain* [New York: Grosset, 1995], and *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* [New York: Harcourt Brace, 1999].) For discussion of this in the context of fictional emotions, see Gendler and Karson Kovakovich, "Genuine Rational Fictional Emotions," in *Contemporary Debates in Aesthetics and the Philosophy of Art*, Matthew Kiernan, ed., (Malden, MA: Blackwell, 2005), pp. 241–53; Paul Harris, *The Work of the Imagination* (Malden, MA: Blackwell, 2000); T. Schroeder and C. Matheson, "Imagination and Emotion" in *The Architecture of the Imagination*, Shaun Nichols, ed., (New York: Oxford, 2006), pp. 19–39.
19. This gives rise to a potential worry: that alief is not a fundamental mental state, but instead an amalgam of several more primitive mental states: those of entertaining content *R*, experiencing affect *A*, and activating behavioral repertoire *B*. I reply: the fact that our current vocabulary requires us to describe alief-content using three separate terms does not show that the state is an amalgam of three others. Indeed, one might even argue that it is out of these more primitive association patterns ('Mama, warmth and comfort, purse lips to drink') that the less fundamental differentiated attitudes like belief, desire, and imagination are constructed. These are cognitive attitudes that rely on the notion of *representation* (and misrepresentation), a distinction between seeming and being, one that is largely absent from the more primitive state of alief. I discuss this issue further in Gendler, "Alief in Action (and Reaction)." (Thanks to Andy Egan for raising this concern.)
20. William James calls the principle that 'the mere act of thinking about a behavior increase[s] the tendency to engage in that behavior' the *principle of ideomotor action*. He writes: 'We may then lay it down for certain that every [mental] representation of a movement awakens in some degree the actual movement which is its object; and awakens it in a maximum degree whenever it is not kept from so doing by an antagonistic representation present simultaneously to the mind' (James, *The Principles of Psychology* [1890], available online at http://psychclassics.yorku.ca/James/Principles/). Or again: 'Merely thinking about a behavior makes it more likely to occur, even if it is unintended . . . the mere act of thinking about a response, even when the thought involved is meant to help prevent the response, has the automatic effect of increasing the likelihood of that response' (John Bargh, Mark Chen, and Lara Burrows, "The Automaticity of Social Behavior," *Journal of Personality and Social*

Psychology LXXI, no. 2 [August 1996]: pp. 230–44, see p. 232, discussing work by Daniel Wegner).

21. As Bargh, Chen, and Burrows write: 'Recent research has shown that attitudes and other affective reactions can be triggered automatically by the mere presence of relevant objects and events . . . without conscious attention or awareness . . .. [They] then exert their influence on thought and behavior' (ibid., p. 230, citations omitted).
22. For discussion of this distinction in the case of belief, see Price, op. cit.; David M. Armstrong, *Belief, Truth and Knowledge* (New York: Cambridge, 1973); William G. Lycan, "Tacit Belief," in *Belief: Form, Content, and Function*, R. J. Bogdan, ed., (New York: Oxford University Press, 1986), pp. 61–82; John R. Searle, *The Rediscovery of the Mind* (Cambridge: MIT Press, 1992); and Robert Audi, "Dispositional Beliefs and Dispositions to Believe," *Noûs* XXVIII (1994): pp. 419–34. (References thanks to Schwitzgebel, op. cit.)
23. Obviously, there need to be some restrictions on what this causal relation looks like: the connection must be nondeviant, and the encounter must not in itself bring the dispositional alief into existence.
24. In fact it is likely that you right now—prompted by the associations set into play through imagining such a case—occurrently alieve something with similar (though decidedly milder) content.
25. For additional discussion, see Gendler, "Alief in Action (and Reaction)."
26. He writes: 'Regarding-as-true [is] . . . involved in . . . believing . . . [in] supposing or assuming, and in propositional imagining as well . . . . To imagine that *p* is to regard *p* as describing how things are . . .. Imagining is therefore a way of regarding a proposition as true—or, to introduce a term, a way of accepting a proposition' (Velleman, "On the Aim of Belief," p. 250). Note that Velleman's use of the term 'acceptance' is somewhat different than that of L. Jonathan Cohen (*An Essay on Belief and Acceptance* [New York: Oxford University Press, 1992]) and Michael Bratman ("Practical Reasoning and Acceptance in a Context," as reprinted in Bratman, *Faces of Intention* [New York: Cambridge University Press, 1999], pp. 15–34).
27. I am here skating over the difficult question of whether there is a uniform rule for stating what one (loosely) alieves when one (strictly) alieves *R-A-B*.
28. Rozin, Maureen Markwith, and Bonnie Ross, "The Sympathetic Magical Law of Similarity, Nominal Realism, and Neglect of Negatives in Response to Negative Labels," *Psychological Science* I, no. 6 (November 1990): pp. 383–84; see p. 383, reporting results from Rozin and Carol J. Nemeroff, "The Laws of Sympathetic Magic: A Psychological Analysis of Similarity and Contagion," in *Cultural Psychology: Essays on Comparative Human Development*, J. Stigler, G. Herdt, and R.A. Schweder, eds., (New York: Cambridge University Press, 1990), pp. 205–32.
29. As Rozin reports, subjects 'knew this response was foolish, but felt the reluctance anyway. This suggests a 'low-level' gut feeling, that can influence behavior in spite of countering cognitions'—"The Sympathetic Magical Law of Similarity, Nominal Realism, and Neglect of Negatives in Response to Negative Labels," p. 383.
30. As in the following joke. Jean-Paul Sartre was sitting in a cafe when a waitress approached him: 'Can I get you something to drink, Monsieur Sartre?' Sartre replied, 'Yes, I'd like a cup of coffee with sugar, but no cream.' Nodding agreement, the waitress walked off to fill the order, returning a few minutes later. 'I'm sorry, Monsieur Sartre,' she said, 'we are all out of cream—would you like your coffee with no milk instead?' (Taken with slight variation from http://www.workjoke.com/projoke70.htm.)
31. Note that they are not hyperopaque in a stronger sense: they *do* permit *salva veritate* substitution of expressions with respect to which the subject holds corresponding patterns of alief. (Thanks to Dave Chalmers for pointing out this stronger reading.)
32. Likewise (in a slight variation on a Kantian theme), my triskaidekaphobia may be elicited by '13' but not by '7+6.' This feature of alief will turn out to be important in the discussion in section IV below.
33. As William James writes: 'To make our nervous system our ally instead of our enemy . . . we must make automatic and habitual, as early as possible, as many useful actions as we can' (op. cit.).
34. It is the reality-sensitivity of belief that is typically taken to explain the impossibility of believing at will. Cf. Bernard Williams, "Deciding to Believe," reprinted with new pagination in Williams, *Problems of the Self* (New York: Cambridge, 1970/1973), pp. 136–51, for a classic articulation of this view. (Thanks to Ted Sider for suggesting that I consider this issue in the context of alief.)
35. Cf. Williams (op. cit.); for instructions, see Pascal, for example, "The Wager," from *Pensées*, reprinted in *The Elements of Philosophy: Readings from Past and Present*, Susanna Siegel Gendler and Steven M. Cahn, eds., (New York: Oxford University Press, 2007).
36. For traditional discussions in addition to Williams (op. cit.), see Barbara Winters, "Believing at Will," *The Journal of Philosophy* LXXVI, no. 5 (May 1979): pp. 243–56; Jonathan Bennett, "Why Is Belief Involuntary?" *Analysis* l, no. 2 (March 1990): pp. 87–107; cf. J. T. Cook, "Deciding to Believe without Self-Deception," *The Journal of Philosophy* LXXXIV, no. 8 (August 1987): pp. 441–46. There has been a recent resurgence of interest in these issues: see, for example, Philip Pettit and Michael Smith, "Freedom in Belief and Desire," *The Journal of Philosophy* XCIII, no. 9 (September 1996): pp. 429–49; Paul Noordhof, "Believe What You Want," *Proceedings of the Aristotelian Society* CI (2001): pp. 247–65; Matthias Steup, "Doxastic Freedom," *Synthese* CLXI, no. 3 (2008): pp. 375–92; and essays cited therein.
37. Velleman (op. cit., p. 255). Velleman rejects this view, for reasons related to the ones discussed here, but notes that the view has been widely endorsed, by philosophers as diverse as R. B. Braithwaite, "The Nature of Believing," *Proceedings of the Aristotelian Society* XXXIII (1932–33): pp. 129–46; David Armstrong, *Belief, Truth, and Knowledge* (New York: Cambridge University Press, 1973); W. V. Quine and J. S. Ullian, *The Web of Belief* (New York: Random House, 1978), 2nd ed.; Robert Stalnaker, *Inquiry* (Cambridge, MA: MIT Press, 1984); Lynn Rudder Baker, *Explaining Attitudes: A Practical Approach to the Mind* (Cambridge: Cambridge University Press, 1995); and Daniel Dennett, "Intentional Systems," *The Journal of Philosophy* LXVIII, no. 4 (February 25, 1971): pp. 87–106, and "Do Animals Have Beliefs?" in

Comparative Approaches to Cognitive Sciences, Herbert Roitblat, ed., (Cambridge, MA: MIT Press, 1995), pp. 111–18.

38. Stalnaker, op. cit., p. 15; cf. Dennett, op. cit.
39. If you are worried about verbal reports counting as behavior, add the requisite caveat that they never speak about this particular belief.
40. See Bratman, "Practical Reasoning and Acceptance in a Context"; Currie, "XI—Imagination as Motivation," *Proceedings of the Aristotelian Society* CII, no. 1 (2002): pp. 201–16; Velleman, "The Aim of Belief"; Egan and Doggett, "Wanting Things You Don't Want," *Philosophers' Imprint*, 7:1-17 (2007). For my own take on these issues, see "On the Relation between Pretense and Belief," in *Imagination, Philosophy, and the Arts*, Matthew Kieran and Dominic McIver Lopes, eds., (New York: Routledge, 2003), pp. 125–41; "Imaginative Contagion," *Metaphilosophy* XXXVII, no. 2 (April 2006): pp. 183–203; and "Self-Deception as Pretense." op. cit.
41. Of course, I may become accustomed to performing the alief-averse action, and my hesitation may dissipate. But this is a way of *changing* alief (by creating new patterns of representational-affective-behavioral association patterns)—not a way of 'throwing it off.'
42. A nice recent defense of such an account can be found in Eric Funkhouser and Shannon Spaulding's "Imagination and Other Scripts," where they defend what they call the "Belief-Desire Thesis: For every intentional action, there is a belief-desire pair that both causes and rationalizes that intentional action" *Philosophical Studies* 143 (3):291-314 (2009).
43. I discuss these and related cases in greater detail in "Imaginative Contagion" (op. cit.); some of the material in this section draws on the discussion in that essay. In the earlier paper, I suggested that these cases were examples of a phenomenon that I called 'imaginative contagion.' I now think that the phenomenon that I identified there is a special case of an alief-like phenomenon. Readers interested in additional examples of these sorts of cases may find them in that essay, and in the works cited therein.
44. I am gliding over many important distinctions about exactly which sorts of primes tend to generate which sorts of responses: whether they tend to elicit assimilation or contrast, whether they involve goals or nongoals, and so forth. In a full-fledged account of alief, it will be important to address these subtleties in proper detail.
45. In such a task, subjects are presented with a list containing a number of five-word sets, and asked to come up with a sentence for each set that contains at least four of the designated words. So, for example, one such set might contain the words 'snow, roof, cat, cheerful, red' and the subject might write: 'The cat stood in the snow atop the red roof.' For original presentation of the scrambled sentence task, see Thomas K. Srull and Robert S. Wyer, "The Role of Category Accessibility in the Interpretation of Information about Persons," *Journal of Personality and Social Psychology* XXXVII, no, 10 (1979): pp. 1660–72, and "Category Accessibility and Social Perception," *Journal of Personality and Social Psychology* XXXVIII, no. 6 (1980): pp. 841–56. For discussion of priming, see (concept) L. H. Storms, "Apparent Backward Association: A Situational Effect," *Journal of Experimental Psychology* LV, no. 4 (1958): pp. 390–95; (term) S. J. Segal and C. N. Cofer, "The Effect of Recency and Recall on Word Association," *The American Psychologist* XV (1960): p. 451; (discussion) J. H. Neely, "Semantic Priming and Retrieval from Lexical Memory," *Journal of Experimental Psychology: General* CVI (1997): pp. 226–54; (review) J. P. Toth and E. M. Reingold, "Beyond Perception: Conceptual Contributions to Unconscious Influences of Memory," in *Implicit Cognition*, G. Underwood, ed., (New York: Oxford University Press, 1996), pp. 41–84.
46. Bargh, Chen, and Burrows, "The Automaticity of Social Behavior," op. cit., p. 236.
47. For either there is no fact of the matter whether interruption in such circumstances is socially acceptable (in which case there is no truth for the mechanisms to be sensitive to), or there is a fact of the matter, which is either independent of or dependent on the subject's attitudes in the situation. If it is independent of those, then the belief-forming mechanism is clearly truth-insensitive, for the three groups using the same mechanism respond in three different ways to the same scenario. (See next note.) And if it is dependent on those attitudes—say: interrupting is socially unacceptable if the interrupter takes it to be socially unacceptable—then the belief-desire explanation to which we are appealing becomes close to vacuous. (This is not to deny that there are all sort of interesting instances of self-fulfilling beliefs and assessment-dependent attitudes. But subliminal primes altering perceptions of rudeness are hardly instances of the *cogito*.)
48. Actually, there is experimental evidence suggesting that the behavior is not the result of any sort of conscious process. 'To assess whether the priming manipulation had resulted in different perception of the experimenter's politeness, Bargh, Chen, and Burrows examined the ratings participants made on' a scale where they were explicitly asked to rate the experimenter's degree of politeness. They found 'no reliable difference in the ratings made in the three priming conditions'—all three groups ranked him as neither especially polite nor especially impolite. They continue: the 'fact that the behavioral measure showed quite strong effects of the priming manipulation, whereas the effect on the judgment measurement was nonexistent, argues against the . . . interpretation . . . that the priming manipulation affected consciously made judgments about the experimenter, which then determined behavioral responses to him. The results instead point to a direct effect on behavior that is not mediated by conscious perceptual or judgment processes' (Bargh, Chen, and Burrows, op. cit., p. 235).
49. See Bargh, Chen, and Burrows (op. cit.). Of course, subjects in the first (neutral) condition also have various aliefs rendered occurrent, but none that systematically affects the likelihood of their interrupting the experimenter; they are like the visually induced aliefs associated with stepping onto the back porch (as opposed to the Skywalk); they are present, but we do not need to appeal to them to explain otherwise discordant behavior.
50. For discussion of how these results can be reconciled with neuropsychological evidence suggesting that simple motor actions are impervious to high-level mental processes such as stereotype activation, see Jane F. Banfield, Louise F. Pendry, Avril J.

Mewse, and Martin G. Edwards, "The Effects of an Elderly Stereotype Prime on Reaching and Grasping Actions," *Social Cognition* XXI, no. 4 (August 2003): pp. 299–319.

51. Though one that I tacitly appealed to in my discussion of this case in "Imaginative Contagion" (op. cit.).
52. In this particular case, 'inspection of the responses' to a similar priming task 'revealed that only 1 of the 19 participants showed any awareness of the relationship between the stimulus words and the elderly stereotype' (Bargh, Chen, and Burrows, op. cit., p. 237).
53. Henk Aarts and Ap Dijksterhuis, "The Silence of the Library," *Journal of Personality and Social Psychology* LXXXIV, no. 1 (January 2003): pp. 18–28.
54. Bargh, Chen, and Burrows, op. cit.
55. G. Fitzsimmons and J. A. Bargh, "Thinking of You," *Journal of Personality and Social Psychology* LXXXIV (2003): pp. 148–63.
56. Polly Shulman, "Priming the Mind," *Science: Science Careers* (March 2007). In addition to showing greater enjoyment of embarrassing media and less emotional distress from violent media, distant-dot drawers offered lower estimations of calories in unhealthy food, and weaker reports of emotional attachments to family members. (See Lawrence Williams and Bargh, "Keeping One's Distance: The Effect of Spatial Distance Cues on Affect and Evaluation," *Psychological Science* XIX, no. 3 [2007]: pp. 302–08.)
57. I discuss this issue in more detail in "Philosophical Thought Experiments, Intuitions and Cognitive Equilibrium," *Midwest Studies in Philosophy* XXXI, no. 1 (2007): pp. 68–89. The discussion in the next three paragraphs draws on the discussion from the opening pages of that paper.
58. David Hume, *A Treatise on Human Nature*, L. A. Selby-Bigge, ed., (New York: Oxford University Press, 1978), II.iii.6.4.
59. In the paper on thought experiments, I go on to explore how this phenomenon might help explain both the effectiveness and the limitations of philosophical thought experiments.
60. Hume, op. cit., "Abstract," pp. 661–62.
61. For exploration of this connection in a related context, see the final paragraph of J. Thomas Cook, "Deciding to Believe without Self-Deception" *The Journal of Philosophy*, no. 8 (August 1987): pp. 441–46; cf. also Myles Burnyeat, "Aristotle on Learning to be Good," in *Essays on Aristotle's Ethics*, Amélie O. Rorty, ed., (Berkeley: University of California Press, 1980), pp. 66–92, as well as Bill Pollard, "Explaining Actions with Habits," *American Philosophical Quarterly* XL, no. 1 (2006): pp. 56–69.
62. Aristotle, *Nicomachean Ethics*, J. L. Ackrill, J. O. Urmson, and David Ross, eds., (New York: Oxford University Press, 1998), pp. 1103–04. Somewhat simplistically, one might say that Aristotelian ethics is an ethics of alief, whereas Kantian ethics is an ethics of belief. I hope to explore this issue in more detail in further work.

C. Internalism, Externalism, and Embodiment

The Meaning of 'Meaning'

Hilary Putnam

Meaning and Extension

. . . Since the Middle Ages at least, writers on the theory of meaning have purported to discover an ambiguity in the ordinary concept of meaning, and have introduced a pair of terms—*extension* and *intension*, or *Sinn* and *Bedeutung*, or whatever—to disambiguate the notion. The *extension* of a term, in customary logical parlance, is simply the set of things the term is true of. Thus, 'rabbit,' in its most common English sense, is true of all and only rabbits, so the extension of 'rabbit' is precisely the set of rabbits. Even this notion—and it is the *least* problematical notion in this cloudy subject—has its problems, however. Apart from problems it inherits from its parent notion of *truth*, the foregoing example of 'rabbit' *in its most common English sense* illustrates one such problem: strictly speaking, it is not a term, but an ordered pair consisting of a term and a 'sense' (or an occasion of use, or something else that distinguishes a term in one sense from the same term used in a different sense) that has an extension. Another problem is this: a 'set,' in the mathematical sense, is a 'yes-no' object; any given object either definitely belongs to *S* or definitely does not belong to *S*, if *S* is a set. But words in a natural language are not generally 'yes-no': there are things of which the description 'tree' is clearly true and things of which the description 'tree' is clearly false, to be sure, but there are a host of borderline cases. Worse, the line between the clear cases and the borderline cases is itself fuzzy. Thus the idealization involved in the notion of *extension*—the idealization involved in supposing that there is such a thing as the set of things of which the term 'tree' is true—is actually very severe.

Recently some mathematicians have investigated the notion of a *fuzzy set*— that is, of an object to which other things belong or do not belong with a given probability or to a given degree, rather than belong 'yes-no.' If one really wanted to formalize the notion of extension as applied to terms in a natural language, it would be necessary to employ 'fuzzy sets' or something similar rather than sets in the classical sense.

The problem of a word's having more than one sense is standardly handled by treating each of the senses as a different word (or rather, by treating the word as if it carried invisible subscripts, thus: 'rabbit$_1$'—animal of a certain kind; 'rabbit$_2$'—coward; and as if 'rabbit$_1$' and 'rabbit$_2$' or whatever were different words entirely). This again involves two very severe idealizations (at least two, that is): supposing that words have discretely many senses, and supposing that the entire repertoire of senses is fixed once and for all. Paul Ziff has recently investigated the extent to which both of these suppositions distort the actual situation in natural language;[1] nevertheless, we will continue to make these idealizations here.

Now consider the compound terms 'creature with a heart' and 'creature with a kidney.' Assuming that every creature with a heart possesses a kidney and vice versa, the extension of these two terms is exactly the same. But they obviously differ in meaning. Supposing that there is a sense of 'meaning' in which meaning = extension, there must be another sense of 'meaning' in which the meaning of a

Excerpted from K. Gunderson, ed., *Language, Mind, and Knowledge* (University of Minnesota Press, 1975), pp. 131–93, with permission of the publisher.

term is not its extension but something else, say the 'concept' associated with the term. Let us call this 'something else' the *intension* of the term. The concept of a creature with a heart is clearly a different concept from the concept of a creature with a kidney. Thus the two terms have different intension. When we say they have different 'meaning,' meaning = intension.

Intension and Extension

Something like the preceding paragraph appears in every standard exposition of the notions 'intension' and 'extension.' But it is not at all satisfactory. Why it is not satisfactory is, in a sense, the burden of this entire essay. But some points can be made at the very outset: first of all, what evidence is there that 'extension' is a sense of the word 'meaning'? The canonical explanation of the notions 'intension' and 'extension' is very much like 'in one sense 'meaning' means *extension* and in the other sense 'meaning' means *meaning.*' The fact is that while the notion of 'extension' is made quite precise, relative to the fundamental logical notion of *truth* (and under the severe idealizations remarked above), the notion of intension is made no more precise than the vague (and, as we shall see, misleading) notion 'concept.' It is as if someone explained the notion 'probability' by saying: 'in one sense 'probability' means frequency, and in the other sense it means *propensity.*' 'Probability' *never* means 'frequency,' and 'propensity' is at least as unclear as 'probability.'

Unclear as it is, the traditional doctrine that the notion 'meaning' possesses the extension/intension ambiguity has certain typical consequences. Most traditional philosophers thought of concepts as something *mental.* Thus the doctrine that the meaning of a term (the meaning 'in the sense of intension,' that is) is a concept carried the implication that meanings are mental entities. Frege and more recently Carnap and his followers, however, rebelled against this 'psychologism,' as they termed it. Feeling that meanings are *public* property—that the *same* meaning can be 'grasped' by more than one person and by persons at different times—they identified concepts (and hence 'intensions' or meanings) with abstract entities rather than mental entities. However, 'grasping' these abstract entities was still an individual psychological act. None of these philosophers doubted that understanding a word (knowing its intension) was just a matter of being in a certain psychological state (somewhat in the way in which knowing how to factor numbers in one's head is just a matter of being in a certain very complex psychological state).

Second, the timeworn example of the two terms 'creature with a kidney' and 'creature with a heart' does show that two terms can have the same extension and yet differ in intension. But it was taken to be obvious that the reverse is impossible: two terms cannot differ in extension and have the same intension. Interestingly, no argument for this impossibility was ever offered. Probably it reflects the tradition of the ancient and medieval philosophers who assumed that the concept corresponding to a term was just a conjunction of predicates, and hence that the concept corresponding to a term must *always* provide a necessary and sufficient condition for falling into the extension of the term.[2] For philosophers like Carnap, who accepted the verifiability theory of meaning, the concept corresponding to a term provided (in the ideal case, where the term had 'complete meaning') a *criterion* for belonging to the extension (not just in the sense of 'necessary and sufficient condition,' but in the strong sense of *way of recognizing* if a given thing falls into the extension or not). Thus these positivistic philosophers were perfectly happy to retain the traditional view on this point. So, theory of meaning came to rest on two unchallenged assumptions:

I. That knowing the meaning of a term is just a matter of being in a certain psychological state (in the sense of 'psychological state,' in which states of memory and psychological dispositions are 'psychological states'; no one thought that knowing the meaning of a word was a continuous state of consciousness, of course).
II. That the meaning of a term (in the sense of 'intension') determines its extension (in the sense that sameness of intension entails sameness of extension).

I shall argue that these two assumptions are not jointly satisfied by *any* notion, let alone any notion of meaning. The traditional concept of meaning is a concept which rests on a false theory.

'Psychological State' and Methodological Solipsism

In order to show this, we need first to clarify the traditional notion of a psychological state. In one sense a state is simply a two-place predicate

whose arguments are an individual and a time. In this sense, *being 5 feet tall, being in pain, knowing the alphabet*, and even *being a thousand miles from Paris* are all states. (Note that the *time* is usually left implicit or 'contextual'; the full form of an atomic sentence of these predicates would be 'x *is five feet tall at time* t,' 'x *is in pain at time* t,' etc.) In science, however, it is customary to restrict the term state to properties which are defined in terms of the parameters of the individual which are fundamental from the point of view of the given science. Thus, being five feet tall is a state (from the point of view of physics); being in pain is a state (from the point of view of mentalistic psychology, at least); knowing the alphabet might be a state (from the point of view of cognitive psychology), although it is hard to say; but being a thousand miles from Paris would *not* naturally be called a *state*. In one sense, a psychological state is simply a state which is studied or described by psychology. In this sense it may be trivially true that, say *knowing the meaning of the word 'water'* is a 'psychological state' (viewed from the standpoint of cognitive psychology). But this is not the sense of psychological state that is at issue in the above assumption (I).

When traditional philosophers talked about psychological states (or 'mental' states), they made an assumption which we may call the assumption of methodological solipsism. This assumption is the assumption that no psychological state, properly so called, presupposes the existence of any individual other than the subject to whom that state is ascribed. (In fact, the assumption was that no psychological state presupposes the existence of the subject's *body* even: if *P* is a psychological state, properly so called, then it must be logically possible for a 'disembodied mind' to be in *P*.) This assumption is pretty explicit in Descartes, but it is implicit in just about the whole of traditional philosophical psychology. Making this assumption is, of course, adopting a *restrictive program*—a program which deliberately limits the scope and nature of psychology to fit certain mentalistic preconceptions or, in some cases, to fit an idealistic reconstruction of knowledge and the world. Just *how* restrictive the program is, however, often goes unnoticed. Such common or garden variety psychological states as *being jealous* have to be reconstructed, for example, if the assumption of methodological solipsism is retained. For, in its ordinary use, *x is jealous of y* entails that *y* exists, and *x is jealous of y's regard for z* entails that both *y* and *z* exist (as well as *x*, of course). Thus *being jealous* and *being jealous of someone's regard for someone else* are not psychological states permitted by the assumption of methodological solipsism. (We shall call them 'psychological states in the wide sense' and refer to the states which are permitted by methodological solipsism as 'psychological states in the narrow sense.') The reconstruction required by methodological solipsism would be to reconstrue *jealousy* so that I can be jealous of my own hallucinations, or of figments of my imagination, etc. Only if we assume that psychological states in the narrow sense have a significant degree of causal closure (so that restricting ourselves to psychological states in the narrow sense will facilitate the statement of psychological *laws*) is there any point in engaging in this reconstruction, or in making the assumption of methodological solipsism. But the three centuries of failure of mentalistic psychology is tremendous evidence against this procedure, in my opinion.

Be that as it may, we can now state more precisely what we claimed at the end of the preceding section. Let *A* and *B* be any two terms which differ in extension. By assumption (II) they must differ in meaning (in the sense of 'intension'). By assumption (I), *knowing the meaning of A* and *knowing the meaning of B* are psychological states *in the narrow sense*—for this is how we shall construe assumption (I). *But these psychological states must determine the extension of the terms A and B just as much as the meanings ('intensions') do.*

To see this, let us try assuming the opposite. Of course, there cannot be two terms *A* and *B* such that *knowing the meaning of A* is the same state as *knowing the meaning of B* even though *A* and *B* have different extensions. For *knowing the meaning of A* isn't just 'grasping the intension' of *A*, whatever that may come to; it is also knowing that the 'intension' that one has 'grasped' *is* the intension of *A*. Thus, someone who knows the meaning of 'wheel' presumably 'grasps the intension' of its German synonym *Rad;* but if he doesn't know that the 'intension' in question is the intension of Rad he isn't said to 'know the meaning of Rad.' If *A* and *B* are different terms, then *knowing the meaning of A* is a different state from *knowing the meaning of B* whether the meanings of *A* and *B* be themselves the same or different. But by the same argument, if I_1 and I_2 are different *intensions* and *A* is a term, then *knowing that* I_1 *is the meaning of A* is a different psychological state

from *knowing that* I_2 *is the meaning of A*. Thus, there cannot be two different logically possible worlds L_1 and L_2 such that, say, Oscar is in the *same* psychological state (in the narrow sense) in L_1 and in L_2 (in all respects), but in L_1 Oscar understands *A* as having the meaning I_1 and in L_2 Oscar understands *A* as having the meaning I_2. (For, if there were, then in L_1 Oscar would be in the psychological state *knowing that* I_1 *is the meaning of A* and in L_2 Oscar would be in the psychological state *knowing that* I_2 *is the meaning of A*, and these are different and even—assuming that *A* has just *one* meaning for Oscar in each world—incompatible psychological states in the narrow sense.)

In short, if *S* is the sort of psychological state we have been discussing—a psychological state of the form *knowing that* I *is the meaning of A*, where *I* is an 'intension' and *A* is a term—then the *same* necessary and sufficient condition for falling into the extension of *A* 'works' in *every* logically possible world in which the speaker is in the psychological state *S*. For the state *S determines* the intension *I*, and by assumption (II) the intension amounts to a necessary and sufficient condition for membership in the *extension.*

If our interpretation of the traditional doctrine of intension and extension is fair to Frege and Carnap, then the whole psychologism/Platonism issue appears somewhat a tempest in a teapot, as far as meaning-theory is concerned. (Of course, it is a very important issue as far as general philosophy of mathematics is concerned.) For even if meanings are 'Platonic' entities rather than 'mental' entities on the Frege–Carnap view, 'grasping' those entities is presumably a psychological state (in the narrow sense). Moreover, the psychological state uniquely determines the 'Platonic' entity. So whether one takes the 'Platonic' entity or the psychological state as the 'meaning' would appear to be somewhat a matter of convention. And taking the psychological state to be the meaning would hardly have the consequence that Frege feared, that meanings would cease to be public. For psychological states are 'public' in the sense that different people (and even people in different epochs) can be in the *same* psychological state. Indeed, Frege's argument against psychologism is only an argument against identifying concepts with mental particulars, not with mental entities in general.

The 'public' character of psychological states entails, in particular, that if Oscar and Elmer understand a word *A* differently, then they must be in different psychological states. For the state of *knowing the intension of A to be, say, I* is the *same* state whether Oscar or Elmer be in it. Thus two speakers cannot be in the same psychological state in all respects and understand the term *A* differently; the psychological state of the speaker determines the intension (and hence, by assumption (II), the extension) of *A*.

It is this last consequence of the joint assumptions (I), (II) that we claim to be false. We claim that it is possible for two speakers to be in exactly the *same* psychological state (in the narrow sense), even though the extension of the term *A* in the idiolect of the one is different from the extension of the term *A* in the idiolect of the other. Extension is not determined by psychological state.

This will be shown in detail in later sections. If this is right, then there are two courses open to one who wants to rescue at least one of the traditional assumptions; to give up the idea that psychological state (in the narrow sense) determines *intension*, or to give up the idea that intension determines extension. We shall consider these alternatives later.

Are Meanings in the Head?

That psychological state does not determine extension will now be shown with the aid of a little science-fiction. For the purpose of the following science-fiction examples, we shall suppose that somewhere in the galaxy there is a planet we shall call Twin Earth. Twin Earth is very much like Earth; in fact, people on Twin Earth even speak *English.* In fact, apart from the differences we shall specify in our science-fiction examples, the reader may suppose that Twin Earth is *exactly* like Earth. He may even suppose that he has a *Doppelgänger*—an identical copy—on Twin Earth, if he wishes, although my stories will not depend on this.

Although some of the people on Twin Earth (say, the ones who call themselves 'Americans' and the ones who call themselves 'Canadians' and the ones who call themselves 'Englishmen,' etc.) speak English, there are, not surprisingly, a few tiny differences which we will now describe between the dialects of English spoken on Twin Earth and Standard English. These differences themselves depend on some of the peculiarities of Twin Earth.

One of the peculiarities of Twin Earth is that the liquid called 'water' is not H_2O but a different liquid whose chemical formula is very

long and complicated. I shall abbreviate this chemical formula simply as *XYZ*. I shall suppose that *XYZ* is indistinguishable from water at normal temperatures and pressures. In particular, it tastes like water and it quenches thirst like water. Also, I shall suppose that the oceans and lakes and seas of Twin Earth contain *XYZ* and not water, that it rains *XYZ* on Twin Earth and not water, etc.

If a spaceship from Earth ever visits Twin Earth, then the supposition at first will be that 'water' has the same meaning on Earth and on Twin Earth. This supposition will be corrected when it is discovered that 'water' on Twin Earth is *XYZ*, and the Earthian spaceship will report somewhat as follows:

'On Twin Earth the word 'water' means *XYZ*.'

(It is this sort of use of the word 'means' which accounts for the doctrine that extension is one sense of 'meaning,' by the way. But note that although 'means' does mean something like *has as extension* in this example, one would *not* say

'On Twin Earth the meaning of the word 'water' is *XYZ*.'

unless, possibly, the fact that 'water is *XYZ*' was known to every adult speaker of English on Twin Earth. We can account for this in terms of the theory of meaning we develop below; for the moment we just remark that although the verb 'means' sometimes means 'has as extension,' the nominalization 'meaning' *never* means 'extension.')

Symmetrically, if a spaceship from Twin Earth ever visits Earth, then the supposition at first will be that the word 'water' has the same meaning on Twin Earth and on Earth. This supposition will be corrected when it is discovered that 'water' on Earth is H_2O, and the Twin Earthian spaceship will report

'On Earth[3] the word 'water' means H_2O.'

Note that there is no problem about the extension of the term 'water.' The word simply has two different meanings (as we say) in the sense in which it is used on Twin Earth, the sense of $water_{TE}$, what *we* call 'water' simply isn't water; while in the sense in which it is used on Earth, the sense of $water_E$, what the Twin Earthians call 'water' simply isn't water. The extension of 'water' in the sense of $water_E$ is the set of all wholes consisting of H_2O molecules, or something like that; the extension of water in the sense of $water_{TE}$ is the set of all wholes consisting of *XYZ* molecules, or something like that.

Now let us roll the time back to about 1750. At that time chemistry was not developed on either Earth or Twin Earth. The typical Earthian speaker of English did not know water consisted of hydrogen and oxygen, and the typical Twin Earthian speaker of English did not know 'water' consisted of *XYZ*. Let $Oscar_1$ be such a typical Earthian English speaker, and let $Oscar_2$ be his counterpart on Twin Earth. You may suppose that there is no belief that $Oscar_1$ had about water that $Oscar_2$ did not have about 'water.' If you like, you may even suppose that $Oscar_1$ and $Oscar_2$ were exact duplicates in appearance, feelings, thoughts, interior monologue, etc. Yet the extension of the term 'water' was just as much H_2O on Earth in 1750 as in 1950; and the extension of the term 'water' was just as much *XYZ* on Twin Earth in 1750 as in 1950. $Oscar_1$ and $Oscar_2$ understood the term 'water' differently in 1750 *although they were in the same psychological state*, and although, given the state of science at the time, it would have taken their scientific communities about fifty years to discover that they understood the term 'water' differently. Thus the extension of the term 'water' (and, in fact, its 'meaning' in the intuitive preanalytical usage of that term) is *not* a function of the psychological state of the speaker by itself.

But, it might be objected, why should we accept it that the term 'water' has the same extension in 1750 and in 1950 (on both Earths)? The logic of natural-kind terms like 'water' is a complicated matter, but the following is a sketch of an answer. Suppose I point to a glass of water and say 'this liquid is called water' (or 'this is called water,' if the marker 'liquid' is clear from the context). My 'ostensive definition' of water has the following empirical presupposition that the body of liquid I am pointing to bears a certain sameness relation (say, *x is the same liquid as y*, or *x is the* $same_L$ as y) to most of the stuff I and other speakers in my linguistic community have on other occasions called 'water.' If this presupposition is false because, say, I am without knowing it pointing to a glass of gin and not a glass of water, then I do not intend my ostensive definition to be accepted. Thus the ostensive definition conveys what might be called a defeasible necessary and sufficient condition: the necessary and sufficient condition for being water is bearing the relation $same_L$ to the stuff in the glass; but this is the necessary and sufficient condition only if the

empirical presupposition is satisfied. If it is not satisfied, then one of a series of, so to speak, 'fallback' conditions becomes activated.

The key point is that the relation $same_L$ is a *theoretical* relation whether something is or is not the same liquid as *this* may take an indeterminate amount of scientific investigation to determine. Moreover, even if a 'definite' answer has been obtained either through scientific investigation or through the application of some 'common sense' test, the answer is *defeasible:* future investigation might reverse even the most 'certain' example. Thus, the fact that an English speaker in 1750 might have called *XYZ* 'water,' while he or his successors would not have called *XYZ* water in 1800 or 1850 does not mean that the 'meaning' of 'water' changed for the average speaker in the interval. In 1750 or in 1850 or in 1950 one might have pointed to, say, the liquid in Lake Michigan as an example of 'water.' What changed was that in 1750 we would have mistakenly thought that *XYZ* bore the relation $same_L$ to the liquid in Lake Michigan, while in 1800 or 1850 we would have known that it did not (I am ignoring the fact that the liquid in Lake Michigan was only dubiously water in 1950, of course).

Let us now modify our science-fiction story. I do not know whether one can make pots and pans out of molybdenum; and if one can make them out of molybdenum, I don't know whether they could be distinguished easily from aluminum pots and pans (I don't know any of this even though I have acquired the word 'molybdenum.') So I shall suppose that molybdenum pots and pans *can't* be distinguished from aluminum pots and pans save by an expert. (To emphasize the point, I repeat that this could be true for all I know, and *a fortiori* it could be true for all I know by virtue of 'knowing the meaning' of the words *aluminum* and *molybdenum.*) We will now suppose that molybdenum is as common on Twin Earth as aluminum is on Earth, and that aluminum is as rare on Twin Earth as molybdenum is on Earth. In particular, we shall assume that 'aluminum' pots and pans are made of molybdenum on Twin Earth. Finally, we shall assume that the words 'aluminum' and 'molybdenum' are *switched* on Twin Earth: 'aluminum' is the name of *molybdenum* and 'molybdenum' is the name of *aluminum.*

This example shares some features with the previous one. If a spaceship from Earth visited Twin Earth, the visitors from Earth probably would not suspect that the 'aluminum' pots and pans on Twin Earth were not made of aluminum, especially when the Twin Earthians *said* they were. But there is one important difference between the two cases. An Earthian metallurgist could tell very easily that 'aluminum' was molybdenum, and a Twin Earthian metallurgist could tell equally easily that aluminum was 'molybdenum.' (The shudder quotes in the preceding sentence indicate Twin Earthian usages.) Whereas in 1750 no one on either Earth or Twin Earth could have distinguished water from 'water,' the confusion of aluminum with 'aluminum' involves only a part of the linguistic communities involved.

The example makes the same point as the preceding one. If $Oscar_1$ and $Oscar_2$ are standard speakers of Earthian English and Twin Earthian English respectively, and neither is chemically or metallurgically sophisticated, then there may be no difference at all in their psychological state when they use the word 'aluminum'; nevertheless we have to say that 'aluminum' has the extension *aluminum* in the idiolect of $Oscar_1$ and the extension *molybdenum* in the idiolect of $Oscar_2$. (Also we have to say that $Oscar_1$ and $Oscar_2$ mean different things by 'aluminum,' that 'aluminum' has a different meaning on Earth than it does on Twin Earth, etc.) Again we see that the psychological state of the speaker does *not* determine the extension (*or* the 'meaning,' speaking preanalytically) of the word.

Before discussing this example further, let me introduce a *non*-science-fiction example. Suppose you are like me and cannot tell an elm from a beech tree. We still say that the extension of 'elm' in my idiolect is the same as the extension of 'elm' in anyone else's, viz., the set of all elm trees, and that the set of all beech trees is the extension of 'beech' in *both* of our idiolects. Thus 'elm' in my idiolect has a different extension from 'beech' in your idiolect (as it should). Is it really credible that this difference in extension is brought about by some difference in our *concepts?* My concept of an elm tree is exactly the same as my concept of a beech tree (I blush to confess). (This shows that the identification of meaning 'in the sense of intension' with *concept* cannot be correct, by the way.) If someone heroically attempts to maintain that the difference between the extension of 'elm' and the extension of 'beech' in *my* idiolect is explained by a difference in my psychological state, then we can always refute him by constructing a 'Twin Earth' example—just let the words 'elm' and 'beech' be switched on Twin Earth (the way 'aluminum' and 'molybdenum' were in the previous example). Moreover, I suppose I have a

Doppelgänger on Twin Earth who is molecule for molecule 'identical' with me (in the sense in which two neckties can be 'identical'). If you are a dualist, then also suppose my *Doppelgänger* thinks the same verbalized thoughts I do, has the same sense data, the same dispositions, etc. It is absurd to think *his* psychological state is one bit different from mine: yet he 'means' *beech* when he says 'elm' and *I* 'mean' *elm* when I say elm. Cut the pie any way you like, 'meanings' just ain't in the *head!*

A Sociolinguistic Hypothesis

The last two examples depend upon a fact about language that seems, surprisingly, never to have been pointed out: that there is *division of linguistic labor.* We could hardly use such words as 'elm' and 'aluminum' if no one possessed a way of recognizing elm trees and aluminum metal; but not everyone to whom the distinction is important has to be able to make the distinction. Let us shift the example: consider *gold.* Gold is important for many reasons: it is a precious metal, it is a monetary metal, it has symbolic value (it is important to most people that the 'gold' wedding ring they wear *really* consist of gold and not just *look* gold), etc. Consider our community as a 'factory': in this 'factory' some people have the 'job' of *wearing gold wedding rings*, other people have the 'job' of *selling gold wedding rings*, still other people have the 'job' of *telling whether or not something is really gold.* It is not at all necessary or efficient that everyone who wears a gold ring (or a gold cufflink, etc.), or discusses the 'gold standard,' etc., engage in buying and selling gold. Nor is it necessary or efficient that everyone who buys and sells gold be able to tell whether or not something is really gold in a society where this form of dishonesty is uncommon (selling fake gold) and in which one can easily consult an expert in case of doubt. And it is *certainly* not necessary or efficient that everyone who has occasion to buy or wear gold be able to tell with any reliability whether or not something is really gold.

The foregoing facts are just examples of mundane division of labor (in a wide sense). But they engender a division of linguistic labor: everyone to whom gold is important for any reason has to *acquire* the word 'gold'; but he does not have to acquire the *method of recognizing* if something is or is not gold. He can rely on a special subclass of speakers. The features that are generally thought to be present in connection with a general name—necessary and sufficient conditions for membership in the extension, ways of recognizing if something is in the extension ('criteria'), etc.—are all present in the linguistic community *considered as a collective body;* but that collective body divides the 'labor' of knowing and employing these various parts of the 'meaning' of 'gold.'

This division of linguistic labor rests upon and presupposes the division of *non*linguistic labor, of course. If only the people who know how to tell if some metal is really gold or not have any reason to have the word 'gold' in their vocabulary, then the word 'gold' will be as the word 'water' was in 1750 with respect to that subclass of speakers, and the other speakers just won't acquire it at all. And some words do not exhibit any division of linguistic labor: 'chair,' for example. But with the increase of division of labor in the society and the rise of science, more and more words begin to exhibit this kind of division of labor. 'Water,' for example, did not exhibit it at all prior to the rise of chemistry. Today it is obviously necessary for every speaker to be able to recognize water (reliably under normal conditions), and probably every adult speaker even knows the necessary and sufficient condition 'water is H_2O,' but only a few adult speakers could distinguish water from liquids which superficially resembled water. In case of doubt, other speakers would rely on the judgement of these 'expert' speakers. Thus the way of recognizing possessed by these 'expert' speakers is also, through them, possessed by the collective linguistic body, even though it is not possessed by each individual member of the body, and in this way the most recherché fact about water may become part of the *social* meaning of the word while being unknown to almost all speakers who acquire the word.

It seems to me that this phenomenon of division of linguistic labor is one which it will be very important for sociolinguistics to investigate. In connection with it, I should like to propose the following hypothesis:

> HYPOTHESIS OF THE UNIVERSALITY OF THE DIVISION OF LINGUISTIC LABOR: Every linguistic community exemplifies the sort of division of linguistic labor just described: that is, possesses at least some terms whose associated 'criteria' are known only to a subset of the speakers who acquire the terms, and whose use by the other speakers depends upon a structured cooperation between them and the speakers in the relevant subsets.

It would be of interest, in particular, to discover if extremely primitive peoples were sometimes exceptions to this hypothesis (which would indicate that the division of linguistic labor is a product of social evolution), or if even they exhibit it. In the latter case, one might conjecture that division of labor, including linguistic labor, is a fundamental trait of our species.

It is easy to see how this phenomenon accounts for some of the examples given above of the failure of the assumptions (I), (II). Whenever a term is subject to the division of linguistic labor, the 'average' speaker who acquires it does not acquire anything that fixes its extension. In particular, his individual psychological state *certainly* does not fix its extension; it is only the sociolinguistic state of the collective linguistic body to which the speaker belongs that fixes the extension.

We may summarize this discussion by pointing out that there are two sorts of tools in the world: there are tools like a hammer or a screwdriver which can be used by one person; and there are tools like a steamship which require the cooperative activity of a number of persons to use. Words have been thought of too much on the model of the first sort of tool.

Indexicality and Rigidity[4]

The first of our science-fiction examples—'water' on Earth and on Twin Earth in 1750—does not involve division of linguistic labor, or at least does not involve it in the same way the examples of 'aluminum' and 'elm' do. There were not (in our story, anyway) any 'experts' on water on Earth in 1750, nor any experts on 'water' on Twin Earth. (The example *can* be construed as involving division of labor *across time*, however. I shall not develop this method of treating the example here.) The example *does* involve things which are of fundamental importance to the theory of reference and also to the theory of necessary truth, which we shall now discuss.

There are two obvious ways of telling someone what one means by a natural-kind term such as 'water' or 'tiger' or 'lemon.' One can give him a so-called ostensive definition—'this (liquid) is water'; 'this (animal) is a tiger'; 'this (fruit) is a lemon'; where the parentheses are meant to indicate that the 'markers' *liquid, animal, fruit*, may be either explicit or implicit. Or one can give him a *description*. In the latter case the description one gives typically consists of one or more markers together with a *stereotype*[5]—a standardized description of features of the kind that are typical, or 'normal,' or at any rate stereotypical. The central features of the stereotype generally are *criteria*—features which in normal situations constitute ways of recognizing if a thing belongs to the kind or, at least, necessary conditions (or probabilistic necessary conditions) for membership in the kind. Not all criteria used by the linguistic community as a collective body are included in the stereotype, and in some cases the stereotypes may be quite weak. Thus (unless I am a very atypical speaker), the stereotype of an elm is just that of a common deciduous tree. These features are indeed necessary conditions for membership in the kind (I mean 'necessary' in a loose sense; I don't think 'elm trees are deciduous' is *analytic*), but they fall far short of constituting a way of recognizing elms. On the other hand, the stereotype of a tiger does enable one to recognize tigers (unless they are albino, or some other atypical circumstance is present), and the stereotype of a lemon generally enables one to recognize lemons. In the extreme case, the stereotype may be *just* the marker: the stereotype of molybdenum might be *just* that molybdenum is a *metal*. Let us consider both of these ways of introducing a term into someone's vocabulary.

Suppose I point to a glass of liquid and say '*this* is water,' in order to teach someone the word 'water.' We have already described some of the empirical presuppositions of this act, and the way in which this kind of meaning-explanation is defeasible. Let us now try to clarify further how it is supposed to be taken.

In what follows, we shall take the notion of 'possible world' as primitive. We do this because we feel that in several senses the notion makes sense and is scientifically important even if it needs to be made more precise. We shall assume further that in at least some cases it is possible to speak of the same individual as existing in more than one possible world.[6] Our discussion leans heavily on the work of Saul Kripke, although the conclusions were obtained independently.

Let W_1 and W_2 be two possible worlds in which I exist and in which this glass exists and in which I am giving a meaning explanation by pointing to this glass and saying 'this is water.' (We do *not* assume that the *liquid* in the glass is the same in both worlds.) Let us suppose that in W_1 the glass is full of H_2O and in W_2 the glass is full of *XYZ*. We shall also suppose that W_1 is the actual world and that *XYZ* is the stuff

typically called 'water' in the world W_2 (so that the relation between English speakers in W_1 and English speakers in W_2 is exactly the same as the relation between English speakers on Earth and English speakers on Twin Earth). Then there are two theories one might have concerning the meaning of 'water':

1. One might hold that 'water' was *world-relative* but *constant* in meaning (i.e., the word has a *constant relative meaning*). In this theory, 'water' *means the same* in W_1 and W_2; it's just that water is H_2O in W_1 and water is *XYZ* in W_2.
2. One might hold that water is H_2O in all worlds (the stuff called 'water' in W_2 isn't water), but 'water' doesn't have the same meaning in W_1 and W_2.

If what was said before about the Twin Earth case was correct, then (2) is clearly the correct theory. When I say '*this* (liquid) is water,' the 'this' is, so to speak, a *de re* 'this'—i.e., the force of my explanation is that 'water' is whatever bears a certain equivalence relation (the relation we called 'same$_L$' above) to the piece of liquid referred to as 'this' *in the actual world.*

We might symbolize the difference between the two theories as a 'scope' difference in the following way. In theory (1), the following is true:

(1′) (For every world *W*) (For every *x* in *W*) (*x* is water ≡ *x* bears same$_L$ to the entity referred to as 'this' in *W*)

while on theory (2):

(2′) (For every world *W*) (For every *x* in *W*) (*x* is water ≡ *x* bears same$_L$ to the entity referred to as 'this' *in the actual world* W_1).

(I call this a 'scope' difference because in (1′) 'the entity referred to as 'this' ' is within the scope of 'For every world *W*'—as the qualifying phrase 'in *W*' makes explicit, whereas in (2′) 'the entity referred to as 'this' ' means 'the entity referred to as 'this' *in the actual world*,' and has thus a reference *independent* of the bound variable '*W*.')

Kripke calls a designator 'rigid' (in a given sentence) if (in that sentence) it refers to the same individual in every possible world in which the designator designates. If we extend the notion of rigidity to substance names, then we may express Kripke's theory and mine by saying that the term 'water' is *rigid.*

The rigidity of the term 'water' follows from the fact that when I give the ostensive definition '*this* (liquid) is water' I intend (2′) and not (1′).

We may also say, following Kripke, that when I give the ostensive definition 'this (liquid) is water,' the demonstrative 'this' is *rigid.*

What Kripke was the first to observe is that this theory of the meaning (or 'use,' or whatever) of the word 'water' (and other natural-kind terms as well) has startling consequences for the theory of necessary truth.

To explain this, let me introduce the notion of a *cross-world relation.* A two-term relation *R* will be called *cross-world* when it is understood in such a way that its extension is a set of ordered pairs of individuals *not all in the same possible world.* For example, it is easy to understand the relation *same height as* as a cross-world relation: just understand it so that, e.g., if *x* is an individual in a world W_1 who is five feet tall (in W_1) and *y* is an individual in W_2 who is five feet tall (in W_2), then the ordered pair *x, y* belongs to the extension of *same height as.* (Since an individual may have different heights in different possible worlds in which that same individual exists, strictly speaking it is not the ordered pair *x, y* that constitutes an element of the extension of *same height as*, but rather the ordered pair *x-in-world-*W_1, y-in-world-W_2.)

Similarly, we can understand the relation *same*$_L$ (same liquid as) as a cross-world relation by understanding it so that a liquid in world W_1 which has the same important physical properties (in W_1) that a liquid in W_2 possesses (in W_2) bears *same*$_L$ to the latter liquid.

Then the theory we have been presenting may be summarized by saying that an entity *x*, in an arbitrary possible world, is *water* if and only if it bears the relation *same*$_L$ (construed as a cross-world relation) to the stuff *we* call 'water' in the *actual* world.

Suppose, now, that I have not yet discovered what the important physical properties of water are (in the actual world)—i.e., I don't yet know that water is H_2O. I may have ways of *recognizing* water that are successful (of course, I may make a small number of mistakes that I won't be able to detect until a later stage in our scientific development) but not know the microstructure of water. If I agree that a liquid with the superficial properties of 'water' but a different microstructure *isn't really water*, then my ways of recognizing water (my 'operational definition,' so to speak) cannot be regarded as an analytical specification of *what it is to be* water. Rather, the operational definition, like the ostensive one, is simply a way of pointing out a standard—pointing out the stuff *in the actual world* such that for *x* to be water, in *any* world,

is for x to bear the relation $same_L$ to the *normal* members of the class of *local* entities that satisfy the operational definition. 'Water' on Twin Earth is not water, even if it satisfies the operational definition, because it doesn't bear $same_L$ to the *local* stuff that satisfies the operational definition, and local stuff that satisfies the operational definition but has a microstructure different from the rest of the local stuff that satisfies the operational definition isn't water either, because it doesn't bear $same_L$ to the *normal* examples of the local 'water.'

Suppose, now, that I discover the microstructure of water—that water is H_2O. At this point I will be able to say that the stuff on Twin Earth that I earlier *mistook* for water isn't really water. In the same way if you describe not another planet in the actual universe, but another possible universe in which there is stuff with the chemical formula *XYZ* which passes the 'operational test' for *water*, we shall have to say that that stuff isn't water but merely *XYZ*. You will not have described a possible world in which 'water is *XYZ*,' but merely a possible world in which there are lakes of *XYZ*, people drink *XYZ* (and not water), or whatever. In fact, once we have discovered the nature of water, nothing counts as a possible world in which water doesn't have that nature. Once we have discovered that water (in the actual world) is H_2O, *nothing counts as a possible world in which water isn't* H_2O. In particular, if a 'logically possible' statement is one that holds in some 'logically possible world,' *it isn't logically possible that water isn't* H_2O.

On the other hand, we can perfectly well imagine having experiences that would convince us (and that would make it rational to believe that) water *isn't* H_2O. In that sense, it is conceivable that water isn't H_2O. It is conceivable but it isn't logically possible! Conceivability is no proof of logical possibility.

Kripke refers to statements which are rationally unrevisable (assuming there are such) as *epistemically necessary*. Statements which are true in all possible worlds he refers to simply as necessary (or sometimes as 'metaphysically necessary'). In this terminology, the point just made can be restated as: a statement can be (metaphysically) necessary and epistemically contingent. Human intuition has no privileged access to metaphysical necessity.

Since Kant there has been a big split between philosophers who thought that all necessary truths were analytic and philosophers who thought that some necessary truths were synthetic a priori. But none of these philosophers thought that a (metaphysically) necessary truth could fail to be a priori: the Kantian tradition was as guilty as the empiricist tradition of equating metaphysical and epistemic necessity. In this sense Kripke's challenge to received doctrine goes far beyond the usual empiricism/Kantianism oscillation.

In this paper our interest is in theory of meaning, however, and not in theory of necessary truth. Points closely related to Kripke's have been made in terms of the notion of *indexicality*.[7] Words like 'now,' 'this,' 'here,' have long been recognized to be *indexical*, or *token-reflexive*—i.e., to have an extension which varied from context to context or token to token. For these words no one has ever suggested the traditional theory that 'intension determines extension.' To take our Twin Earth example: if I have a *Doppelgänger* on Twin Earth, then when I think 'I have a headache,' *he* thinks 'I have a headache.' But the extension of the particular token of 'I' in his verbalized thought is himself (or his unit class, to be precise), while the extension of the token of 'I' in *my* verbalized thought is *me* (or my unit class, to be precise). So the same word, 'I,' has two different extensions in two different idiolects; but it does not follow that the concept I have of myself is in any way different from the concept my *Doppelgänger* has of himself.

Now then, we have maintained that indexicality extends beyond the *obviously* indexical words and morphemes (e.g., the tenses of verbs). Our theory can be summarized as saying that words like 'water' have an unnoticed indexical component: 'water' is stuff that bears a certain similarity relation to the water *around here*. Water at another time or in another place or even in another possible world has to bear the relation $same_L$ to *our* 'water' *in order to be water*. Thus the theory that (1) words have 'intensions,' which are something like concepts associated with the words by speakers; and that (2) intension determines extension—cannot be true of natural-kind words like 'water' for the same reason the theory cannot be true of obviously indexical words like 'I.'

The theory that natural-kind words like 'water' are indexical leaves it open, however, whether to say that 'water' in the Twin Earth dialect of English has the same *meaning* as 'water' in the Earth dialect and a different extension (which is what we normally say about 'I' in different idiolects), thereby giving up the doctrine that 'meaning (intension) determines extension'; or to say, as we have chosen to do,

that difference in extension is *ipso facto* a difference in meaning for natural-kind words, thereby giving up the doctrine that meanings are concepts, or, indeed, mental entities of *any* kind.

It should be clear, however, that Kripke's doctrine that natural-kind words are rigid designators and our doctrine that they are indexical are but two ways of making the same point. We heartily endorse what Kripke says when he writes:

> Let us suppose that we do fix the reference of a name by a description. Even if we do so, we do not then make the name synonymous with the description, but instead we use the name rigidly to refer to the object so named, even in talking about counterfactual situations where the thing named would not satisfy the description in question. Now, this is what I think is in fact true for those cases of naming where the reference is fixed by description. But, in fact, I also think, contrary to most recent theorists, that the reference of names is rarely or almost never fixed by means of description. And by this I do not just mean what Searle says: 'It's not a single description, but rather a cluster, a family of properties that fixes the reference.' I mean that properties in this sense are not used at all.[8]

Other Words

. . . So far we have only used natural-kind words as examples, but the points we have made apply to many other kinds of words as well. They apply to the great majority of all nouns, and to other parts of speech as well.

Let us consider for a moment the names of artifacts—words like 'pencil,' 'chair,' 'bottle,' etc. The traditional view is that these words are certainly defined by conjunctions, or possibly clusters, of properties. Anything with all of the properties in the conjunction (or sufficiently many of the properties in the cluster, on the cluster model) is necessarily a *pencil, chair, bottle*, or whatever. In addition, some of the properties in the cluster (on the cluster model) are usually held to be *necessary* (on the conjunction-of-properties model, *all* of the properties in the conjunction are necessary). *Being an artifact* is supposedly necessary, and belonging to a kind with a certain standard purpose—e.g., 'pencils are artifacts,' and 'pencils are standardly intended to be written with' are supposed to be necessary. Finally, this sort of necessity is held to be *epistemic* necessity—in fact, analyticity.

Let us once again engage in science fiction. This time we use an example devised by Rogers Albritton. Imagine that we someday discover that *pencils are organisms.* We cut them open and examine them under the electron microscope, and we see the almost invisible tracery of nerves and other organs. We spy upon them, and we see them spawn, and we see the offspring grow into full-grown pencils. We discover that these organisms are not imitating other (artifactual) pencils—there are not and never were any pencils except these organisms. It is strange, to be sure, that there is *lettering* on many of these organisms—e.g., BONDED *Grants* DELUXE made in U.S.A. No. 2.—perhaps they are intelligent organisms, and this is their form of camouflage. (We also have to explain why no one ever attempted to manufacture pencils, etc., but this is clearly a possible world, in some sense.)

If this is conceivable, and I agree with Albritton that it is, then it is epistemically possible that *pencils could turn out to be organisms.* It follows that *pencils are artifacts* is not epistemically necessary in the strongest sense and, a fortiori, not analytic.

Let us be careful, however. Have we shown that there is a possible world in which pencils are organisms? I think not. What we have shown is that there is a possible world in which certain organisms are the *epistemic counterparts* of pencils (the phrase is Kripke's). To return to the device of Twin Earth: imagine this time that pencils on Earth are just what we think they are, artifacts manufactured to be written with, while 'pencils' on Twin Earth are organisms à la Albritton. Imagine, further, that this is totally unsuspected by the Twin Earthians—they have exactly the beliefs about 'pencils' that we have about pencils. When we discovered this, we would not say: 'some pencils are organisms.' We would be far more likely to say: 'the things on Twin Earth that pass for pencils aren't really pencils. They're really a species of organism.'

Suppose now the situation to be as in Albritton's example both on Earth and on Twin Earth. Then we would say 'pencils are organisms.' Thus, whether the 'pencil-organisms' on Twin Earth (or in another possible universe) are really *pencils* or not is a function of whether or not the *local* pencils are organisms or not. If the local pencils are just what we think they are, then a possible world in which there are pencilorganisms is *not* a possible world in which *pencils are organisms;* there are *no* possible worlds in which pencils are organisms in this case (which is, of course, the actual one). That pencils are artifacts *is* necessary in the sense of true in all

possible worlds—metaphysically necessary. But it doesn't follow that it's epistemically necessary.

It follows that 'pencil' is not *synonymous* with any description—not even loosely synonymous with a *loose* description. When we use the word 'pencil,' we intend to refer to whatever has the same *nature* as the normal examples of the local pencils in the actual world. 'Pencil' is just as *indexical* as 'water' or 'gold.'

In a way, the case of pencils turning out to be organisms is complementary to the case we discussed some years ago[9] of cats turning out to be robots (remotely controlled from Mars). Katz[10] argues that we misdescribed this case: that the case should rather be described as its *turning out that there are no cats in this world.* Katz admits that we might *say* 'Cats have turned out not to be animals, but robots'; but he argues that this is a semantically deviant sentence which is glossed as 'the things I am referring to as 'cats' have turned out not to be animals, but robots.' Katz's theory is bad linguistics, however. First of all, the explanation of how it is we can *say* 'Cats are robots' is simply an all-purpose explanation of how we can say *anything.* More important, Katz's theory predicts that 'Cats are robots' is *deviant*, while 'There are no cats in the world' is nondeviant, in fact standard, in the case described. Now then, I don't deny that there *is* a case in which 'There are not (and never were) any cats in the world' would be standard: we might (speaking epistemically) discover that we have been suffering from a collective hallucination. ('Cats' are like pink elephants.) But in the case I described, 'Cats have turned out to be robots remotely controlled from Mars' is surely nondeviant, and 'There are no cats in the world' is highly deviant.

Incidentally, Katz's account is not only bad linguistics; it is also bad as a rational reconstruction. The reason we *don't* use 'cat' as synonymous with a description is surely that we know enough about cats to know that they do have a hidden structure, and it is good scientific methodology to use the name to refer rigidly to the things that possess that hidden structure, and not to whatever happens to satisfy some description. Of course, if we *knew* the hidden structure we could frame a description in terms of *it;* but we don't at this point. In this sense the use of natural-kind words reflects an important fact about our relation to the world: we know that there are kinds of things with common hidden structure, but we don't yet have the knowledge to describe all those hidden structures.

Katz's view has more plausibility in the 'pencil' case than in the 'cat' case, however. We think we *know* a necessary and sufficient condition for being a *pencil*, albeit a vague one. So it is possible to make 'pencil' synonymous with a loose description. We *might* say, in the case that 'pencils turned out to be organisms' *either* 'Pencils have turned out to be organisms' *or* 'There are no pencils in the world'—i.e., we might use 'pencil' either as a natural-kind word or as a 'one-criterion' word.[11]

On the other hand, we might doubt that there *are* any true one-criterion words in natural language, apart from stipulative contexts. Couldn't it turn out that pediatricians aren't doctors but Martian spies? Answer 'yes,' and you have abandoned the synonymy of 'pediatrician' and 'doctor specializing in the care of children.' It seems that there is a strong tendency for words which are introduced as 'one-criterion' words to develop a 'natural-kind' sense, with all the concomitant rigidity and indexicality. In the case of artifact-names, this natural-kind sense seems to be the predominant one.

(There is a joke about a patient who is on the verge of being discharged from an insane asylum. The doctors have been questioning him for some time, and he has been giving perfectly sane responses. They decide to let him leave, and at the end of the interview one of the doctors inquires casually, 'What do you want to be when you get out?' 'A teakettle.' The joke would not be intelligible if it were literally inconceivable that a person could be a teakettle.)

There are, however, words which retain an almost pure one-criterion character. These are words whose meaning derives from a transformation: *hunter = one who hunts.*

Not only does the account given here apply to most nouns, but it also applies to other parts of speech. Verbs like 'grow,' adjectives like 'red,' etc., all have indexical features. On the other hand, some syncategorematic words seem to have more of a one-criterion character. 'Whole,' for example, can be explained thus: *The army surrounded the town* could be true even if the A division did not take part. *The whole army surrounded the town* means every part of the army (of the relevant kind, e.g., the A Division) took part in the action signified by the verb.[12]

Meaning

Let us now see where we are with respect to the notion of meaning. We have now seen that the extension of a term is not fixed by a concept that the individual speaker has in his

head, and this is true both because extension is, in general, determined *socially*—there is division of linguistic labor as much as of 'real' labor—and because extension is, in part, determined *indexically*. The extension of our terms depends upon the actual nature of the particular things that serve as paradigms,[13] and this actual nature is not, in general, fully known to the speaker. Traditional semantic theory leaves out only two contributions to the determination of extension—the contribution of society and the contribution of the real world!

We saw at the outset that meaning cannot be identified with extension. Yet it cannot be identified with 'intension' either, if intension is something like an individual speaker's *concept*. What are we to do?

There are two plausible routes that we might take. One route would be to retain the identification of meaning with concept and pay the price of giving up the idea that meaning determines extension. If we followed this route, we might say that 'water' has the same *meaning* on Earth and on Twin Earth, but a different *extension*. (Not just a different *local* extension but a different *global* extension. The *XYZ* on Twin Earth isn't in the extension of the tokens of 'water' that I utter, but it is in the extension of the tokens of 'water' that my *Doppelgänger* utters, and this isn't just because Twin Earth is far away from me, since molecules of H_2O are in the extension of the tokens of 'water' that I utter no matter how far away from me they are in space and time. Also, what I can counterfactually suppose water to be is different from what my *Doppelgänger* can counterfactually suppose 'water' to be.) While this is the correct route to take for an *absolutely* indexical word like 'I,' it seems incorrect for the words we have been discussing. Consider 'elm' and 'beech,' for example. If these are 'switched' on Twin Earth, then surely we would *not* say that 'elm' has the same meaning on Earth and Twin Earth, even if my *Doppelgänger*'s stereotype of a beech (or an 'elm,' as he calls it) is identical with my stereotype of an elm. Rather, we would say that 'elm' in my *Doppelgänger*'s idiolect means *beech*. For this reason, it seems preferable to take a different route and identify 'meaning' with an ordered pair (or possibly an ordered *n-tuple*) of entities, *one of which is the extension*. (The other components of the, so to speak, 'meaning vector' will be specified later.) Doing this makes it trivially true that *meaning determines extension* (i.e., difference in extension is ipso facto difference in meaning), but totally abandons the idea that if there is a difference in the meaning my *Doppelgänger* and I assign to a word, then there *must* be some difference in our concepts (or in our psychological state). Following this route, we can say that my *Doppelgänger* and *I mean something different* when we say 'elm,' but this will not be an assertion about our psychological states. All this means is that the tokens of the word he utters have a different extension than the tokens of the word I utter; but this difference in extension is not a reflection of any difference in our individual linguistic competence considered in isolation.

If this is correct, and I think it is, then the traditional problem of meaning splits into two problems. The first problem is to account for the *determination of extension*. Since, in many cases, extension is determined socially and not individually, owing to the division of linguistic labor, I believe that this problem is properly a problem for sociolinguistics. Solving it would involve spelling out in detail exactly how the division of linguistic labor works. The so-called 'causal theory of reference,' introduced by Kripke for proper names and extended by us to natural-kind words and physical-magnitude terms, falls into this province. For the fact that, in many contexts, we assign to the tokens of a name that I utter whatever referent we assign to the tokens of the same name uttered by the person from whom I acquired the name (so that the reference is transmitted from speaker to speaker, starting from the speakers who were present at the 'naming ceremony,' even though no fixed *description* is transmitted) is simply a special case of social cooperation in the determination of reference.

The other problem is to describe *individual competence*. Extension may be determined socially, in many cases, but we don't assign the standard extension to the tokens of a word *W* uttered by Jones *no matter how* Jones uses *W*. Jones has to have some particular ideas and skills in connection with *W* in order to play his part in the linguistic division of labor. Once we give up the idea that individual competence has to be so strong as to actually determine extension, we can begin to study it in a fresh frame of mind. . . .

The Meaning of 'Meaning'

We may now summarize what has been said in the form of a proposal concerning how one might reconstruct the notion of 'meaning.' Our

proposal is not the only one that might be advanced on the basis of these ideas, but it may serve to encapsulate some of the major points. In addition, I feel that it recovers as much of ordinary usage in common sense talk and in linguistics as one is likely to be able to conveniently preserve. Since, in my view something like the assumptions (I) and (II) listed in the first part of this paper are deeply embedded in ordinary meaning talk, and these assumptions are jointly inconsistent with the facts, no reconstruction is going to be without some counterintuitive consequences.

Briefly, my proposal is to define 'meaning' not by picking out an object which will be identified with the meaning (although that might be done in the usual set-theoretic style if one insists), but by specifying a normal form (or, rather, a *type* of normal form) for the description of meaning. If we know what a 'normal form description' of the meaning of a word should be, then, as far as I am concerned, we know what meaning is in any scientifically interesting sense.

My proposal is that the normal form description of the meaning of a word should be a finite sequence, or 'vector,' whose components should certainly include the following (it might be desirable to have other types of components as well): (1) the syntactic markers that apply to the word, e.g., 'noun'; (2) the semantic markers that apply to the word, e.g., 'animal,' 'period of time'; (3) a description of the additional features of the stereotype, if any; (4) a description of the extension.

The following convention is a part of this proposal: the components of the vector all represent a hypothesis about the individual speaker's competence, *except the extension.* Thus the normal form description for 'water' might be, in part:

Syntactic Markers
mass noun; concrete;

Semantic Markers
natural-kind; liquid;

Stereotype
colorless; transparent; tasteless; thirst-quenching; etc.

Extension
H_2O (give or take impurities)

—this does not mean that knowledge of the fact that water is H_2O is being imputed to the individual speaker or even to the society. It means that (*we* say) the extension of the term 'water' as *they* (the speakers in question) use it is *in fact* H_2O. The objection 'who are *we* to say what the extension of *their* term is in fact' has been discussed above. Note that this is fundamentally an objection to the notion of *truth*, and that extension is a relative of truth and inherits the family problems.

Let us call two descriptions *equivalent* if they are the same except for the description of the extension, and the two descriptions are coextensive. Then, if the set variously described in the two descriptions is, *in fact*, the extension of the word in question, and the other components in the description are correct characterizations of the various aspects of competence they represent, *both* descriptions count as correct. Equivalent descriptions are both correct or both incorrect. This is another way of making the point that, although we have to use a *description* of the extension to *give* the extension, we think of the component in question as being the *extension* (the *set*), not the description of the extension.

In particular the representation of the words 'water' in Earth dialect and 'water' in Twin Earth dialect would be the same except that in the last column the normal form description of the Twin Earth word 'water' would have *XYZ* and not H_2O. This means, in view of what has just been said, that we are ascribing the *same* linguistic competence to the typical Earthling/Twin Earthian speaker, but a different extension to the word, nonetheless.

This proposal means that we keep assumption (II) of our early discussion. Meaning determines extension—by construction, so to speak. But (I) is given up; the psychological state of the individual speaker does not determine 'what he means.'

In most contexts this will agree with the way we speak, I believe. But one paradox: suppose Oscar is a German-English bilingual. In our view, in his total collection of dialects, the words 'beech' and *Buche* are *exact synonyms.* The normal form descriptions of their meanings would be identical. But he might very well not know that they are synonyms! A speaker can have two synonyms in his vocabulary and not know that they are synonyms!

It is instructive to see how the failure of the apparently obvious 'if S_1 and S_2 are synonyms and Oscar understands both S_1 and S_2 then Oscar knows that S_1 and S_2 are synonyms' is related to the falsity of (I), in our analysis. Notice that if we had chosen to omit the extension as a component

of the 'meaning-vector,' which is David Lewis's proposal as I understand it, then we would have the paradox that 'elm' and 'beech' have the *same meaning* but different extensions!

On just about any materialist theory, believing a proposition is likely to involve processing some *representation* of that proposition, be it a sentence in a language, a piece of 'brain code,' a thought form, or whatever. Materialists, and not only materialists, are reluctant to think that one can believe propositions *neat*. But even materialists tend to believe that, if one believes a proposition, *which* representation one employs is (pardon the pun) immaterial. If S_1 and S_2 are both representations that are *available* to me, then if I believe the proposition expressed by S_1 under the representation S_1, I must also believe it under the representation S_2—at least, I must do this if I have any claim to rationality. But, as we have just seen, this isn't right. Oscar may well believe that *this* is a 'beech' (it has a sign on it that says 'beech'), but not believe or disbelieve that this is a '*Buche*.' It is not just that belief is a process involving representations; he believes the proposition (if one wants to introduce 'propositions' at all) under one representation and not under another.

The amazing thing about the theory of meaning is how long the subject has been in the grip of philosophical misconceptions, and how strong these misconceptions are. Meaning has been identified with a necessary and sufficient condition by philosopher after philosopher. In the empiricist tradition, it has been identified with a method of verification, again by philosopher after philosopher. Nor have these misconceptions had the virtue of exclusiveness; not a few philosophers have held that meaning = method of verification = necessary and sufficient condition.

On the other side, it is amazing how weak the grip of the facts has been. After all, what have been pointed out in this essay are little more than home truths about the way we use words and how much (or rather, how little) we actually know when we use them. My own reflection on these matters began after I published a paper in which I confidently maintained that the meaning of a word was 'a battery of semantical rules,'[14] and then began to wonder how the meaning of the common word 'gold' could be accounted for in this way. And it is not that philosophers had never considered such examples: Locke, for example, uses this word as an example and is not troubled by the idea that its meaning is a necessary and sufficient condition!

If there is a reason for both learned and lay opinion having gone so far astray with respect to a topic which deals, after all, with matters which are in everyone's experience, matters concerning which we all have more data than we know what to do with, matters concerning which we have, if we shed preconceptions, pretty clear intuitions, it must be connected to the fact that the grotesquely mistaken views of language which are and always have been current reflect two specific and very central philosophical tendencies: the tendency to treat cognition as a purely *individual* matter and the tendency to ignore the *world*, insofar as it consists of more than the individual's 'observations.' Ignoring the division of linguistic labor is ignoring the social dimension of cognition; ignoring what we have called the *indexicality* of most words is ignoring the contribution of the environment. Traditional philosophy of language, like much traditional philosophy, leaves out other people and the world; a better philosophy and a better science of language must encompass both.

NOTES

1. This is discussed by Ziff, *Understanding Understanding* (Ithaca, NY: Cornell University Press, 1972), especially chapter VIII.
2. This tradition grew up because *the* term whose analysis provoked all the discussion in medieval philosophy was the term 'God,' and the term 'God' was thought to be defined through the conjunction of the terms 'Good,' 'Powerful,' 'Omniscient,' etc.—the so-called 'Perfections.' There was a problem, however, because God was supposed to be a Unity, and Unity was thought to exclude His essence being complex in *any* way—i.e., 'God' was defined through a conjunction of terms, but God (without quotes) could not be the logical product of properties, nor could He be the unique thing exemplifying the logical product of two or more *distinct* properties, because even this highly abstract kind of 'complexity' was held to be incompatible with His perfection of Unity. This is a theological paradox with which Jewish, Arabic, and Christian theologians wrestled for centuries (e.g., the doctrine of the Negation of Privation in Maimonides and Aquinas). It is amusing that theories of contemporary interest, such as conceptualism and nominalism, were first proposed as solutions to the problem of predication in the case of God. It is also amusing that the favorite model of definition in all of this theology—the conjunction-of-properties model—should survive, at least through its consequences, in philosophy of language until the present day.

3. Rather, they will report: 'On Twin Earth (*the Twin Earthian name for Terra*—H.P.), the word 'water' means H_2O.'
4. The substance of this section was presented at a series of lectures I gave at the University of Washington (Summer Institute in Philosophy) in 1968, and at a lecture at the University of Minnesota.
5. See my "Is Semantics Possible," *Metaphilosophy*, 1, no. 3 (July 1970).
6. This assumption is not actually needed in what follows. What *is* needed is that the same *natural kind* can exist in more than one possible world.
7. These points were made in my 1968 lectures at the University of Washington and the University of Minnesota.
8. See Kripke's "Identity and Necessity," in *Identity and Individuation*, M. Munitz, ed., (New York: New York University Press, 1972), p. 157.
9. See my "It Ain't Necessarily So," *Journal of Philosophy* 59 (1962): pp. 658–71.
10. See Katz, "Logic and Language: An Examination of Recent Criticisms of Intentionalism," in *Language, Mind, and Knowledge*, K. Gunderson, ed., (Minneapolis: University of Minnesota Press, 1975).
11. The idea of a 'one-criterion' word, and a theory of analyticity based on this notion, appears in my "The Analytic and The Synthetic," in *Minnesota Studies in the Philosophy of Science*, vol. 3, H. Feigl and G. Maxwell, eds., (Minneapolis: University of Minnesota Press, 1962).
12. This example comes from an analysis by Anthony Kroch (in his M.I.T. doctoral dissertation, 1974, Department of Linguistics).
13. I *don't* have in mind the Flewish notion of 'paradigm' in which any paradigm of a *K* is *necessarily* a *K* (in reality).
14. "How Not to Talk about Meaning," in *Boston Studies in the Philosophy of Science*, vol. 2, R. Cohen and M. Wortofsky, eds., (New York: Humanities Press, 1965).

18 Individualism and the Mental

Tyler Burge

Since Hegel's *Phenomenology of Spirit*, a broad, inarticulate division of emphasis between the individual and his social environment has marked philosophical discussions of mind. On one hand, there is the traditional concern with the individual subject of mental states and events. In the elderly Cartesian tradition, the spotlight is on what exists or transpires 'in' the individual—his secret cogitations, his innate cognitive structures, his private perceptions and introspections, his grasping of ideas, concepts, or forms. More evidentially oriented movements, such as behaviorism and its liberalized progeny, have highlighted the individual's publicly observable behavior—his input-output relations and the dispositions, states, or events that mediate them. But both Cartesian and behaviorist viewpoints tend to feature the individual subject. On the other hand, there is the Hegelian preoccupation with the role of social institutions in shaping the individual and the content of his thought. This tradition has dominated the continent since Hegel. But it has found echoes in English-speaking philosophy during this century in the form of a concentration on language. Much philosophical work on language and mind has been in the interests of Cartesian or behaviorist viewpoints that I shall term 'individualistic.' But many of Wittgenstein's remarks about mental representation point up a social orientation that is discernible from his flirtations with behaviorism. And more recent work on the theory of reference has provided glimpses of the role of social cooperation in determining what an individual thinks.

In many respects, of course, these emphases within philosophy—individualistic and social—are compatible. To an extent, they may be regarded simply as different currents in the turbulent stream of ideas that has washed the intellectual landscape during the last hundred and some odd years. But the role of the social environment has received considerably less clear-headed philosophical attention (though perhaps not less philosophical attention) than the role of the states, occurrences, or acts in, on,

Excerpted from P. French, T. Uehling, and H. Wettstein, eds., *Studies in Metaphysics* (University of Minnesota Press, 1979), with permission of publisher and author.

or by the individual. Philosophical discussions of social factors have tended to be obscure, evocative, metaphorical, or platitudinous, or to be bent on establishing some large thesis about the course of history and the destiny of man. There remains much room for sharp delineation. I shall offer some considerations that stress social factors in descriptions of an individual's mental phenomena. These considerations call into question individualistic presuppositions of several traditional and modern treatments of mind. I shall conclude with some remarks about mental models.

I. Terminological Matters

Our ordinary mentalistic discourse divides broadly into two sorts of idiom. One typically makes reference to mental states or events in terms of sentential expressions. The other does not. A clear case of the first kind of idiom is 'Alfred thinks that his friends' sofa is ugly.' A clear case of the second sort is 'Alfred is in pain.' Thoughts, beliefs, intentions, and so forth are typically specified in terms of subordinate sentential clauses, that-clauses, which may be judged as true or false. Pains, feels, tickles, and so forth have no special semantical relation to sentences or to truth or falsity. There are intentional idioms that fall in the second category on this characterization, but that share important semantical features with expressions in the first—idioms like 'Al worships Buicks.' But I shall not sort these out here. I shall discuss only the former kind of mentalistic idiom. The extension of the discussion to other intentional idioms will not be difficult.

In an ordinary sense, the noun phrases that embed sentential expressions in mentalistic idioms provide the *content* of the mental state or event. We shall call that-clauses and their grammatical variants '*content-clauses.*' Thus, the expression 'that sofas are more comfortable than pews' provides the content of Alfred's belief that sofas are more comfortable than pews. My phrase 'provides the content' represents an attempt at remaining neutral, at least for present purposes, among various semantical and metaphysical accounts of precisely how that-clauses function and precisely what, if anything, contents are.

Although the notion of content is, for present purposes, ontologically neutral, I do think of it as holding a place in a systematic *theory* of mentalistic language. The question of when to count contents different, and when the same, is answerable to theoretical restrictions. It is often remarked that in a given context we may ascribe to a person two that-clauses that are only loosely equivalent and count them as attributions of the 'same attitude.' We may say that Al's intention to climb Mt. McKinley and his intention to climb the highest mountain in the United States are the 'same intention.' (I intend the terms for the mountain to occur obliquely here. See later discussion.) This sort of point extends even to content clauses with extensionally nonequivalent counterpart notions. For contextually relevant purposes, we might count a thought that the glass contains some water as 'the same thought' as a thought that the glass contains some thirst-quenching liquid, particularly if we have no reason to attribute either content as opposed to the other, and distinctions between them are irrelevant. Nevertheless, in both these examples, every systematic theory I know of would want to represent the semantical contribution of the content-clauses in distinguishable ways—as 'providing different contents.'

One reason for doing so is that the person himself is capable of having different attitudes described by the different content-clauses, even if these differences are irrelevant in a particular context. (Al might have developed the intention to climb the highest mountain before developing the intention to climb Mt. McKinley—regardless of whether he, in fact, did so.) A second reason is that the counterpart components of the that-clauses allude to distinguishable elements in people's cognitive lives. 'Mt. McKinley' and 'the highest mountain in the U.S.' serve, or might serve, to indicate cognitively different notions. This is a vague, informal way of generalizing Frege's point: the thought that Mt. McKinley is the highest mountain in the U.S. is potentially interesting or informative. The thought that Mt. McKinley is Mt. McKinley is not. Thus, when we say in a given context that attribution of different contents is attribution of the 'same attitude,' we use 'same attitude' in a way similar to the way we use 'same car' when we say that people who drive Fords (or green 1970 Ford Mavericks) drive the 'same car.' For contextual purposes different cars are counted as 'amounting to the same.'

Although this use of 'content' is theoretical, it is not, I think, theoretically controversial. In cases where we shall be counting contents different, the cases will be uncontentious: in any systematic theory, differences in the

extension—the actual denotation, referent, or application—of counterpart expressions in that-clauses will be semantically represented, and will, in our terms, make for differences in content. I shall be avoiding the more controversial, but interesting, questions about the general conditions under which sentences in that-clauses can be expected to provide the same content.

I should also warn of some subsidiary terms. I shall be (and have been) using the term '*notion*' to apply to components or elements of contents. Just as whole that-clauses provide the content of a person's attitude, semantically relevant components of that-clauses will be taken to indicate notions that enter into the attitude (or the attitude's content). The term is supposed to be just as ontologically neutral as its fellow. When I talk of understanding or mastering the notion of contract, I am not relying on any special epistemic or ontological theory, except insofar as the earlier-mentioned theoretical restrictions on the notion of content are inherited by the notion of notion. The expression, '*understanding (mastering) a notion*' is to be construed more or less intuitively. Understanding the notion of contract comes roughly to knowing what a contract is. One can master the notion of contract without mastering the term 'contract'—at the very least if one speaks some language other than English that has a term roughly synonymous with 'contract.' (An analogous point holds for my use of 'mastering a content.') Talk of notions is roughly similar to talk of concepts in an informal sense. 'Notion' has the advantage of being easier to separate from traditional theoretical commitments.

I speak of *attributing* an attitude, content, or notion, and of *ascribing* a that-clause or other piece of language. Ascriptions are the linguistic analogs of attributions. This use of 'ascribe' is nonstandard, but convenient and easily assimilated.

There are semantical complexities involving the behavior of expressions in content-clauses, most of which we can skirt. But some must be touched on. Basic to the subject is the observation that expressions in content-clauses are often not intersubstitutable with extensionally equivalent expressions in such a way as to maintain the truth value of the containing sentence. Thus, from the facts that water is H_2O and that Bertrand thought that water is not fit to drink, it does not follow that Bertrand thought that H_2O is not fit to drink. When an expression like 'water' functions in a content-clause so that it is not freely exchangeable with all extensionally equivalent expressions, we shall say that it has *oblique occurrence*. Roughly speaking, the reason why 'water' and 'H_2O' are not interchangeable in our report of Bertrand's thought is that 'water' plays a role in characterizing a different mental act or state from that which 'H_2O' would play a role in characterizing. In this context at least, thinking that water is not fit to drink is different from thinking that H_2O is not fit to drink.

By contrast, there are nonoblique occurrences of expressions in content-clauses. One might say that some water—say, the water in the glass over there—is thought by Bertrand to be impure; or that Bertrand thought that *that* water is impure. And one might intend to make no distinction that would be lost by replacing 'water' with 'H_2O,'—or 'that water' with 'that H_2O' or 'that common liquid,' or any other expression extensionally equivalent with 'that water.' We might allow these exchanges even though Bertrand had never heard of, say, H_2O. In such purely nonoblique occurrences, 'water' plays *no role* in providing the *content* of Bertrand's thought, *on our use of 'content,'* or (in any narrow sense) in characterizing Bertrand or his mental state. Nor is the water part of Bertrand's thought content. We speak of Bertrand *thinking his content of* the water. At its nonoblique occurrence, the term 'that water' simply isolates, in one of many equally good ways, a portion of wet stuff to which Bertrand or his thought is related or applied. In certain cases, it may also mark a context in which Bertrand's thought is applied. But it is expressions at oblique occurrences within content clauses that primarily do the job of providing the content of mental states or events, and in characterizing the person.

Mentalistic discourse containing obliquely occurring expressions has traditionally been called *intentional discourse*. The historical reasons for this nomenclature are complex and partly confused. But roughly speaking, grammatical contexts involving oblique occurrences have been fixed upon as specially relevant to the representational character (sometimes called 'intentionality') of mental states and events. Clearly oblique occurrences in mentalistic discourse have something to do with characterizing a person's epistemic perspective—how things seem to him, or in an informal sense, how they are represented to him. So without endorsing all the commitments of this tradition, I shall take over its terminology.

The crucial point in the preceding discussion is the assumption that obliquely occurring

expressions in content-clauses are a primary means of identifying a person's intentional mental states or events. A further point is worth remarking here. It is normal to suppose that those content clauses correctly ascribable to a person that are not in general intersubstitutable *salva veritate*—and certainly those that involve extensionally nonequivalent counterpart expressions—identify different mental states or events.

I have cited contextual exceptions to this normal supposition, at least in a manner of speaking. We sometimes count distinctions in content irrelevant for purposes of a given attribution, particularly where our evidence for the precise content of a person or animal's attitude is skimpy. Different contents may contextually identify (what amount to) the 'same attitude.' I have indicated that even in these contexts, I think it best, strictly speaking, to construe distinct contents as describing different mental states or events that are merely equivalent for the purposes at hand. I believe that this view is widely accepted. But nothing I say will depend on it. For any distinct contents, there will be imaginable contexts of attribution in which, even in the loosest, most informal ways of speaking, those contents would be said to describe different mental states or events. This is a consequence of the theoretical role of contents, discussed earlier. Since our discussion will have an 'in principle' character, I shall take these contexts to be the relevant ones. Most of the cases we discuss will involve *extensional* differences between obliquely occurring counterpart expressions in that-clauses. In such cases, it is particularly natural and normal to take different contents as identifying different mental states or events.

II. A Thought Experiment

IIa. First Case

We now turn to a three-step thought experiment. Suppose first that:

> A given person has a large number of attitudes commonly attributed with content-clauses containing 'arthritis' in oblique occurrence. For example, he thinks (correctly) that he has had arthritis for years, that his arthritis in his wrists and fingers is more painful than his arthritis in his ankles, that it is better to have arthritis than cancer of the liver, that stiffening joints is a symptom of arthritis, that certain sorts of aches are characteristic of arthritis, that there are various kinds of arthritis, and so forth. In short, he has a wide range of such attitudes. In addition to these unsurprising attitudes, he thinks falsely that he has developed arthritis in the thigh.

Generally competent in English, rational and intelligent, the patient reports to his doctor his fear that his arthritis has now lodged in his thigh. The doctor replies by telling him that this cannot be so, since arthritis is specifically an inflammation of joints. Any dictionary could have told him the same. The patient is surprised, but relinquishes his view and goes on to ask what might be wrong with his thigh.

The second step of the thought experiment consists of a counterfactual supposition. We are to conceive of a situation in which the patient proceeds from birth through the same course of physical events that he actually does, right to and including the time at which he first reports his fear to his doctor. Precisely the same things (nonintentionally described) happen to him. He has the same physiological history, the same diseases, the same internal physical occurrences. He goes through the same motions, engages in the same behavior, has the same sensory intake (physiologically described). His dispositions to respond to stimuli are explained in physical theory as the effects of the same proximate causes. All of this extends to his interaction with linguistic expressions. He says and hears the same words (word forms) at the same time he actually does. He develops the disposition to assent to 'Arthritis can occur in the thigh' and 'I have arthritis in the thigh' as a result of the same physically described proximate causes. Such dispositions might have arisen in a number of ways. But we can suppose that in both actual and counterfactual situations, he acquires the word 'arthritis' from casual conversation or reading, and never hearing anything to prejudice him for or against applying it in the way that he does, he applies the word to an ailment in his thigh (or to ailments in the limbs of others) which seems to produce pains or other symptoms roughly similar to the disease in his hands and ankles. In both actual and counterfactual cases, the disposition is never reinforced or extinguished up until the time when he expresses himself to his doctor. We further imagine that the patient's nonintentional, phenomenal experience is the same. He has the same pains, visual fields, images, and internal verbal rehearsals. The *counterfactuality* in the supposition touches only the patient's social environment. In actual fact, 'arthritis,' as

used in his community, does not apply to ailments outside joints. Indeed, it fails to do so by a standard, nontechnical dictionary definition. But in our imagined case, physicians, lexicographers, and informed laymen apply 'arthritis' not only to arthritis but to various other rheumatoid ailments. The standard use of the term is to be conceived to encompass the patient's actual misuse. We could imagine either that arthritis had not been singled out as a family of diseases, or that some other term besides 'arthritis' were applied, though not commonly by laymen, specifically to arthritis. We may also suppose that this difference and those necessarily associated with it are the only differences between the counterfactual situation and the actual one. (Other people besides the patient will, of course, behave differently.) To summarize the second step:

> The person might have had the same physical history and nonintentional mental phenomena while the word 'arthritis' was conventionally applied, and defined to apply, to various rheumatoid ailments, including the one in the person's thigh, as well as to arthritis.

The final step is an interpretation of the counterfactual case, or an addition to it as so far described. It is reasonable to suppose that:

> In the counterfactual situation, the patient lacks some—probably *all*—of the attitudes commonly attributed with content-clauses containing 'arthritis' in oblique occurrence. He lacks the occurrent thoughts or beliefs that he has arthritis in the thigh, that he has had arthritis for years, that stiffening joints and various sorts of aches are symptoms of arthritis, that his father had arthritis, and so on.

We suppose that in the counterfactual case we cannot correctly ascribe any content-clause containing an oblique occurrence of the term 'arthritis.' It is hard to see how the patient could have picked up the notion of arthritis. The word 'arthritis' in the counterfactual community does not mean *arthritis*. It does not apply only to inflammations of the joints. We suppose that no other word in the patient's repertoire means *arthritis*. 'Arthritis,' in the counterfactual situation, differs both in dictionary definition and in extension from 'arthritis' as we use it. Our ascriptions of content-clauses to the patient (and ascriptions within his community) would not constitute attributions of the same contents we actually attribute. For counterpart expressions in the content clauses that are actually counterfactually ascribable are not even extensionally equivalent. However we describe the patient's attitudes in the counterfactual situation, it will not be with a term or phrase extensionally equivalent with 'arthritis.' So the patient's counterfactual-attitudes contents differ from his actual ones.

The upshot of these reflections is that the patient's mental contents differ while his entire physical and nonintentional mental histories, considered in isolation from their social context, remain the same. (We could have supposed that he dropped dead at the time he first expressed his fear to the doctor.) The differences seem to stem from differences 'outside' the patient considered as an isolated physical organism, causal mechanism, or seat of consciousness. The difference in his mental contents is attributable to differences in his social environment. In sum, the patient's internal qualitative experiences, his physiological states and events, his behaviorally described stimuli and responses, his dispositions to behave, and whatever sequences of states (nonintentionally described) mediated his input and output—all these remain constant, while his attitude contents differ, even in the extensions of counterpart notions. As we observed at the outset, such differences are ordinarily taken to spell differences in mental states and events.

IIb. Further Exemplifications

The argument has an extremely wide application. It does not depend, for example, on the kind of word 'arthritis' is. We could have used an artifact term, an ordinary natural-kind word, a color adjective, a social-role term, a term for a historical style, an abstract noun, an action verb, a physical-movement verb, or any of various other sorts of words. I prefer to leave open precisely how far one can generalize the argument. But I think it has a very wide scope. The argument can get under way in any case where it is intuitively possible to attribute a mental state or event whose content involves a notion that the subject incompletely understands. As will become clear, this possibility is the key to the thought experiment. I want to give a more concrete sense of the possibility before going further.

It is useful to reflect on the number and variety of intuitively clear cases in which it is normal to attribute a content that the subject incompletely understands. One need only thumb through a dictionary for an hour or so to develop a sense of the extent to which one's beliefs are

infected by incomplete understanding.[1] The phenomenon is rampant in our pluralistic age.

(a.) Most cases of incomplete understanding that support the thought experiment will be fairly idiosyncratic. There is a reason for this. Common linguistic errors, if entrenched, tend to become common usage. But a generally competent speaker is bound to have numerous words in his repertoire, possibly even common words, that he somewhat misconstrues. Many of these misconstruals will not be such as to deflect ordinary ascriptions of that-clauses involving the incompletely mastered term in oblique occurrence. For example, one can imagine a generally competent, rational adult having a large number of attitudes involving the notion of sofa—including beliefs that *those* (some sofas) are sofas, that some sofas are beige, that his neighbors have a new sofa, that he would rather sit in a sofa for an hour than on a church pew. In addition, he might think that sufficiently broad (but single-seat) overstuffed armchairs are sofas. With care, one can develop a thought experiment parallel to the one in section IIa, in which at least some of the person's attitude contents (particularly, in this case, contents of occurrent mental events) differ, while his physical history, dispositions to behavior, and phenomenal experience—non-intentionally and asocially described—remain the same.

(b.) Although most relevant misconstruals are fairly idiosyncratic, there do seem to be certain types of error which are relatively common—but not so common and uniform as to suggest that the relevant terms take on new sense. Much of our vocabulary is taken over from others who, being specialists, understand our terms better than we do.[2] The use of scientific terms by laymen is a rich source of cases. As the arthritis example illustrates, the thought experiment does not depend on specially technical terms. I shall leave it to the imagination of the reader to spin out further examples of this sort.

(c.) One need not look to the laymen's acquisitions from science for examples. People used to buying beef brisket in stores or ordering it in restaurants (and conversant with it in a general way) probably often develop mistaken beliefs (or uncertainties) about just what brisket is. For example, one might think that brisket is a cut from the flank or rump, or that it includes not only the lower part of the chest but also the upper part, or that it is specifically a cut of beef and not of, say, pork. No one hesitates to ascribe to such people content-clauses with 'brisket' in oblique occurrence. For example, a person may believe that he is eating brisket under these circumstances (where 'brisket' occurs in oblique position); or he may think that brisket tends to be tougher than loin. Some of these attitudes may be false; many will be true. We can imagine a counterfactual case in which the person's physical history, his dispositions, and his nonintentional mental life, are all the same, but in which 'brisket' is commonly applied in a different way—perhaps in precisely the way the person thinks it applies. For example, it might apply only to beef and to the upper and lower parts of the chest. In such a case, as in the sofa and arthritis cases, it would seem that the person would (or might) lack some or all of the propositional attitudes that are actually attributed with content clauses involving 'brisket' in oblique position.

(d.) Someone only generally versed in music history, or superficially acquainted with a few drawings of musical instruments, might naturally but mistakenly come to think that clavichords included harpsichords without legs. He may have many other beliefs involving the notion of clavichord, and many of these may be true. Again, with some care, a relevant thought experiment can be generated.

(e.) A fairly common mistake among lawyers' clients is to think that one cannot have a contract with someone unless there has been a written agreement. The client might be clear in intending 'contract' (in the relevant sense) to apply to agreements, not to pieces of paper. Yet he may take it as part of the meaning of the word, or the essence of law, that a piece of formal writing is a necessary condition for establishing a contract. His only experiences with contracts might have involved formal documents, and he undergeneralizes. It is not terribly important here whether one says that the client misunderstands the term's meaning, or alternatively that the client makes a mistake about the essence of contracts. In either case, he misconceives what a contract is, yet ascriptions involving the term in oblique position are made anyway.

It is worth emphasizing here that I intend the misconception to involve the subject's attaching counterfactual consequences to his mistaken belief about contracts. Let me elaborate this a bit. A common dictionary definition of 'contract' is 'legally binding agreement.' As I am imagining the case, the client does not explicitly define 'contract' to himself in this way (though he might use this phrase in explicating

the term). And he is not merely making a mistake about what the law happens to enforce. If asked why unwritten agreements are not contracts, he is likely to say something like, 'They just aren't' or 'It is part of the nature of the law and legal practice that they have no force.' He is not disposed without prodding to answer, 'It would be possible but impractical to give unwritten agreements legal force.' He might concede this. But he would add that such agreements would not be contracts. He regards a document as inseparable from contractual obligation, regardless of whether he takes this to be a matter of meaning or a metaphysical essentialist truth about contracts.

Needless to say, these niceties are philosophers' distinctions. They are not something an ordinary man is likely to have strong opinions about. My point is that the thought experiment is independent of these distinctions. It does not depend on misunderstandings of dictionary meaning. One might say that the client understood the term's dictionary meaning, but misunderstood its essential application in the law—misconceived the nature of contracts. The thought experiment still flies. In a counterfactual case in which the law enforces both written and unwritten agreements and in which the subject's behavior and so forth are the same, but in which 'contract' *means* 'legally binding agreement based on written document,' we would not attribute to him a mistaken belief that a contract requires written agreement, although the lawyer might have to point out that there are other legally binding agreements that do not require documents. Similarly, the client's other propositional attitudes would no longer involve the notion of contract, but another more restricted notion.

(f.) People sometimes make mistakes about color ranges. They may correctly apply a color term to a certain color, but also mistakenly apply it to shades of a neighboring color. When asked to explain the color term, they cite the standard cases (for 'red,' the color of blood, fire engines, and so forth). But they apply the term somewhat beyond its conventionally established range—beyond the reach of its vague borders. They think that fire engines, including *that* one, are red. They observe that red roses are covering the trellis. But they also think that *those* things are a shade of red (whereas they are not). Second looks do not change their opinion. But they give in when other speakers confidently correct them in unison.

This case extends the point of the contract example. The error is linguistic or conceptual in something like the way that the shopper's mistake involving the notion of brisket is. It is not an ordinary empirical error. But one may reasonably doubt that the subjects misunderstand the dictionary meaning of the color term. Holding their nonintentional phenomenal experience, physical history, and behavioral dispositions constant, we can imagine that 'red' were applied as they mistakenly apply it. In such cases, we would no longer ascribe content-clauses involving the term 'red' in oblique position. The attribution of the correct beliefs about fire engines and roses would be no less affected than the attribution of the beliefs that, in the actual case, display the misapplication. Cases bearing out the latter point are common in anthropological reports on communities whose color terms do not match ours. Attributions of content typically allow for the differences in conventionally established color ranges.

Here is not the place to refine our rough distinctions among the various kinds of misconceptions that serve the thought experiment. Our philosophical purposes do not depend on how these distinctions are drawn. Still, it is important to see what an array of conceptual errors is common among us. And it is important to note that such errors do not always or automatically prevent attribution of mental content provided by the very terms that are incompletely understood or misapplied. The thought experiment is nourished by this aspect of common practice.

IIc. Expansion and Delineation of the Thought Experiment

As I have tried to suggest in the preceding examples, the relevant attributions in the first step of the thought experiment need not display the subject's error. They may be attributions of a true content. We can begin with a propositional attitude that involved the misconceived notion, but in a true, unproblematic application of it: for example, the patient's belief that he, like his father, developed arthritis in the ankles and wrists at age 58 (where 'arthritis' occurs obliquely).

One need not even rely on an underlying *mis*conception in the thought experiment. One may pick a case in which the subject only partially understands an expression. He may apply it firmly and correctly in a range of cases, but be unclear or agnostic about certain of its applications or implications which, in fact, are fully established in common practice. Most of the examples we gave previously can be reinterpreted

in this way. To take a new one, imagine that our protagonist is unsure whether his father has mortgages on the car and the house, or just one on the house. He is a little uncertain about exactly how the loan and collateral must be arranged in order for there to be a mortgage, and he is not clear about whether one may have mortgages on anything other than houses. He is sure, however, that Uncle Harry paid off his mortgage. Imagine our man constant in the ways previously indicated and that 'mortgage' commonly applied only to mortgages on houses. But imagine banking practices themselves to be the same. Then the subject's uncertainty would plausibly not involve the notion of mortgage. Nor would his other propositional attitudes be correctly attributed with the term 'mortgage' in oblique position. Partial understanding is as good as misunderstanding for our purposes.

On the other hand, the thought experiment does appear to depend on the possibility of someone's having a propositional attitude despite an incomplete mastery of some notion in its content. To see why this appears to be so, let us try to run through a thought experiment, attempting to avoid any imputation of incomplete understanding. Suppose the subject thinks falsely that all swans are white. One can certainly hold the features of swans and the subject's nonintentional phenomenal experience, physical history, and nonintentional dispositions constant, and imagine that 'swan' meant 'white swan' (and perhaps some other term, unfamiliar to the subject, meant what 'swan' means). Could one reasonably interpret the subject as having different attitude contents without at some point invoking a misconception? The questions to be asked here are about the subject's dispositions. For example, in the actual case, if he were shown a black swan and told that he was wrong, would he fairly naturally concede his mistake? Or would he respond, 'I'm doubtful that that's a swan' until we brought in dictionaries, encyclopedias, and other native speakers to correct his usage? In the latter case, his understanding of 'swan' would be deviant. Suppose then that in the actual situation he would respond normally to the counterexample. Then there is reason to say that he understands the notion of swan correctly: and his error is not conceptual or linguistic, but empirical in an ordinary and narrow sense. (Of course, the line we are drawing here is pretty fuzzy.) When one comes to the counterfactual stage of the thought experiment, the subject has the same dispositions to respond pliably to the presentation of a black specimen. But such a response would suggest a misunderstanding of the term 'swan' as counterfactually used. For in the counterfactual community what they call 'swans' could not fail to be white. The mere presentation of a black swan would be irrelevant to the definitional truth 'All swans are white.' I have not set this case up as an example of the thought experiment's going through. Rather I have used it to support the conjecture that *if* the thought experiment is to work, one must at some stage find the subject believing (or having some attitude characterized by) a content, despite an incomplete understanding or misapplication. An ordinary empirical error appears not to be sufficient.

It would be a mistake, however, to think that incomplete understanding, in the sense that the argument requires, is in general an unusual or even deviant phenomenon. *What I have called 'partial understanding' is common or even normal in the case of a large number of expressions in our vocabularies.* 'Arthritis' is a case in point. Even if by the grace of circumstance a person does not fall into views that run counter to the term's meaning or application, it would not be the least deviant or 'socially unacceptable' to have no clear attitude that would block such views. 'Brisket,' 'contract,' 'recession,' 'sonata,' 'deer,' 'elm' (to borrow a well-known example), 'ore-amplifier,' 'carburetor,' 'gothic,' 'fermentation' probably provide analogous cases. Continuing the list is largely a matter of patience. The sort of 'incomplete understanding' required by the thought experiment includes quite ordinary, nondeviant phenomena.

It is worth remarking that the thought experiment as originally presented might be run in reverse. The idea would be to start with an ordinary belief or thought involving no incomplete understanding. Then we find the incomplete understanding in the second step. For example, properly understanding 'arthritis,' a patient may think (correctly) that he has arthritis. He happens to have heard of arthritis only occurring in joints, and he correctly believes that that is where arthritis always occurs. Holding his physical history, dispositions, and pain constant, we imagine that 'arthritis' commonly applies to rheumatoid ailments of all sorts. Arthritis has not been singled out for special mention. If a patient were told by a doctor 'You also have arthritis in the thigh,' the patient would be disposed (as he is in the actual case) to respond, 'Really? I didn't know that one could have arthritis except in joints.'

The doctor would answer, 'No, arthritis occurs in muscles, tendons, bursas, and elsewhere.' The patient would stand corrected. The notion that the doctor and patient would be operating with in such a case would not be that of arthritis.

My reasons for not having originally set out the thought experiment in this way are largely heuristic. As will be seen, discussion of the thought experiment will tend to center on the step involving incomplete understanding. And I wanted to encourage you, dear reader, to imagine actual cases of incomplete understanding in your own linguistic community. Ordinary intuitions in the domestic case are perhaps less subject to premature warping in the interests of theory. Cases involving not only mental-content attribution, but also translation of a foreign tongue, are more vulnerable to intrusion of side issues.

A secondary reason for not beginning with this 'reversed' version of the thought experiment is that I find it doubtful whether the thought experiment always works in symmetric fashion. There may be special intuitive problems in certain cases—perhaps, for example, cases involving perceptual natural kinds. We may give special interpretations to individual misconceptions in imagined foreign communities, when those misconceptions seem to match our conceptions. In other words, there may be some systematic, intuitive bias in favor of at least certain of our notions for purposes of interpreting the misconceptions of imagined foreigners. I do not want to explore the point here. I think that any such bias is not always crucial, and that the thought experiment frequently works 'symmetrically.' We have to take account of a person's community in interpreting his words and describing his attitudes—and this holds in the foreign case as well as in the domestic case.

The reversal of the thought experiment brings home the important point that *even those propositional attitudes not infected by incomplete understanding* depend for their content on social factors that are independent of the individual, asocially and non-intentionally described. For if the social environment has been appropriately different, the contents of those attitudes would have been different.

Even *apart* from reversals of the thought experiment, it is plausible (in the light of its original versions) that our well-understood propositional attitudes depend partly for their content on social factors independent of the individual, asocially and nonintentionally construed. For each of us can reason as follows. Take a set of attitudes that involve a given notion and whose contents are well-understood by me. It is only contingent that I understand that notion as well as I do. Now holding my community's practices constant, imagine that I understand the given notion incompletely, but that the deficient understanding is such that it does not prevent my having attitude contents involving that notion. In fact, imagine that I am in the situation envisaged in the first step of one of the original thought experiments. In such a case, a proper subset of the original set of my actual attitude contents would, or might, remain the same—intuitively, at least those of my actual attitudes whose justification or point is untouched by my imagined deficient understanding. (In the arthritis case, an example would be a true belief that many old people have arthritis.) These attitude contents remain constant despite the fact that my understanding, inference patterns, behavior, dispositions, and so on would in important ways be different and partly inappropriate to applications of the given notion. What is it that enables these unaffected contents to remain applications of the relevant notion? It is not *just* that my understanding, inference patterns, behavior, and so forth are enough like my actual understanding, inference patterns, behavior, and so forth. For if communal practice had *also* varied so as to apply the relevant notion as I am imagining I misapply it, then my attitude contents would not involve the relevant notion at all. This argument suggests that communal practice is a factor (in addition to my understanding, inference patterns, and perhaps behavior, physical activity, and other features) in fixing the contents of my attitudes—even in cases where I fully understand the content.

IId. Independence from Factive-Verb and Indexical-Reference Paradigms

The thought experiment does not play on psychological 'success' verbs or 'factive' verbs—verbs like 'know,' 'regret,' 'realize,' 'remember,' 'foresee,' 'perceive.' This point is important for our purposes because such verbs suggest an easy and clear-cut distinction between the contribution of the individual subject and the object, 'veridical' contribution of the environment to making the verbs applicable. (Actually the matter becomes more complicated on reflection, but we shall stay with the simple cases.) When a person knows that snow

is common in Greenland, his knowledge obviously depends on more than the way the person is; it depends on there actually being a lot of snow in Greenland. His mental state (belief that snow is common in Greenland) must be successful in a certain way (true). By changing the environment, one could change the truth value of the content, so that the subject could no longer be said to know the content. It is part of the burden of our argument that even intentional mental states of the individual like beliefs, which carry no implication of veridicality or success, cannot be understood by focusing purely on the individual's acts, dispositions, and 'inner' goings-on.

The thought experiment also does not rest on the phenomenon of indexicality, or on *de re* attitudes, in any direct way. When Alfred refers to an apple, saying to himself 'That is wholesome,' what he refers to depends not just on the content of what he says or thinks, but on what apple is before him. Without altering the meaning of Alfred's utterance, the nature of his perceptual experiences, or his physical acts or dispositions, we could conceive an exchange of the actual apple for another one that is indistinguishable to Alfred. We would thereby conceive him as referring to something different and even as saying something with a different truth value.

This rather obvious point about indexicality has come to be seen as providing a model for understanding a certain range of mental states or events—*de re* attitudes. The precise characterization of this range is no simple philosophical task. But the clearest cases involve nonobliquely occurring terms in content clauses. When we say that Bertrand thinks of some water that it would not slake his thirst (where 'water' occurs in purely nonoblique position), we attribute a *de re* belief to Bertrand. We assume that Bertrand has something like an indexical relation to the water. The fact that Bertrand believes something of some water, rather than of a portion of some other liquid that is indistinguishable to him, depends partly on the fact that it is water to which Bertrand is contextually, 'indexically' related. For intuitively we could have exchanged the liquids without changing Bertrand and thereby changed what Bertrand believed his belief content *of*—and even whether his belief was true of it.[3] It is easy to interpret such cases by holding that the subject's mental states and contents (with allowances for brute differences in the contexts in which he applies those contents) remain the same. The differences in the situations do not pertain in any fundamental way to the subject's mind or the nature of his mental content, but to how his mind or content is related to the world.

It seems to me clear that the thought experiment need not rely on *de re* attitudes at all. The subject need not have entered into special *en rapport* or quasi-indexical relations with objects that the misunderstood term applies to in order for the argument to work. We can appeal to attitudes that would usually be regarded as paradigmatic cases of *de dicto*, nonindexical, *non-de-re* mental attitudes or events. The primary mistake in the contract example is one such, but we could choose others to suit the reader's taste. To insist that such attitudes must all be indexically infected of *de re* would, I think, be to trivialize and emasculate these notions, making nearly all attitudes *de re*. All *de dicto* attitudes presuppose *de re* attitudes. But it does not follow that indexical or *de re* elements survive in every attitude (cf. notes 2 and 3).

I shall not, however, argue this point here. The claim that is crucial is not that our argument does not fix on *de re* attitudes. It is, rather, that the social differences between the actual and counterfactual situations affect the *content* of the subject's attitudes. That is, the difference affects standard cases of obliquely occurring, cognitive-content-conveying expressions in content-clauses. For example, still with his misunderstanding, the subject might think that this (referring to his disease in his hands) is arthritis. Or he might think *de re* of the disease in his ankle or of the disease in his thigh that his arthritis is painful. It does not really matter whether the relevant attitude is *de re* or purely *de dicto*. What is crucial to our argument is that the occurrence of 'arthritis' is oblique and contributes to a characterization of the subject's mental content. One might even hold, implausibly I think, that all the subject's attitudes involving the notion of arthritis are *de re*, that 'arthritis' in that-clauses *indexically* picks out the property of being arthritis, or something like that. The fact remains that the term occurs obliquely in the relevant cases and serves in characterizing the *dicta* or contents of the subject's attitudes. The thought experiment exploits this fact.

Approaches to the mental that I shall later criticize as excessively individualistic tend to assimilate environmental aspects of mental phenomena to either the factive-verb or indexical-reference paradigm (cf. note 2). This sort of assimilation suggests that one might maintain a relatively clear-cut distinction

between extramental and mental aspects of mentalistic attributions. And it may encourage the idea that the distinctively mental aspects can be understood fundamentally in terms of the individual's abilities, dispositions, states, and so forth, considered in isolation from his social surroundings. Our argument undermines this later suggestion. Social content infects even the distinctively mental features of mentalistic attributions. No man's intentional mental phenomena are insular. Every man is a piece of the social continent, a part of the social main. . . .

NOTES

1. Our examples suggest points about learning that need exploration. It would seem naive to think that we first attain a mastery of expressions or notions we use and then tackle the subject matters we speak and think about in using those expressions or notions. In most cases, the processes overlap. But while the subject's understanding is still partial, we sometimes attribute mental contents in the very terms the subject has yet to master. Traditional views take mastering a word to consist in matching it with an already mastered (or innate) concept. But it would seem, rather, that many concepts (or mental content components) are like words in that they may be employed before they are mastered. In both cases, employment appears to be an integral part of the process of mastery.
2. A development of a similar theme may be found in Hilary Putnam's notion of a division of linguistic labor [cf. "The Meaning of 'Meaning,'" *Philosophical Papers* 2 (London: Cambridge University Press, 1975), pp. 227 ff.]. Putnam's imaginative work is in other ways congenial with points I have developed. Some of his examples can be adapted in fairly obvious ways so as to give an argument with different premises, but a conclusion complementary to the one I arrive at in Section IIa:

   Consider Alfred's belief contents involving the notion of water. Without changing Alfred's (or his fellows') nonintentional phenomenal experiences, internal physical occurrences, or dispositions to respond to stimuli on sensory surfaces, we can imagine that not water (H_2O), but a different liquid with different structure but similar macro-properties (and identical phenomenal properties) played the role in his environment that water does in ours. In such a case, we could ascribe no content clauses to Alfred with 'water' in oblique position. His belief contents would differ. The conclusion (with which I am in sympathy) is that mental contents are affected not only by the physical and qualitatively mental way the person is, but by the nature of his *physical environment.*

   Putnam himself does not give quite this argument. He nowhere states the first and third steps, though he gives analogs of them for the meaning of 'water.' This is partly just a result of his concentration on meaning instead of propositional attitudes. But some of what he says even seems to oppose the argument's conclusion. He remarks in effect that the subject's *thoughts* remain constant between his actual and counterfactual cases (p. 224). In his own argument he explicates the difference between actual and counterfactual cases in terms of a difference in the extension of term, not a difference in those aspects of their meaning that play a role in the cognitive life of the subject. And he tries to explicate his examples in terms of indexicality—a mistake, I think, and one that tends to divert attention from major implications of the examples he gives (cf. Section IId). In my view, the examples do illustrate the fact that all attitudes involving natural-kind notions, including *de dicto* attitudes, presuppose *de re* attitudes. But the examples do not show that natural-kind linguistic expressions are in any ordinary sense indexical. Nor do they show that beliefs involving natural-kind notions are always *de re.* Even if they did, the change from actual to counterfactual cases would affect oblique occurrences of natural-kind terms in that-clauses—occurrences that are the key to attributions of cognitive content (cf. above and note 3). In the cited paper and earlier ones, much of what Putnam says about psychological states (and implies about mental states) has a distinctly individualistic ring. Below in Section IV (not reprinted here—*ed.*), I criticize viewpoints about mental phenomena influenced by and at least strongly suggested in his earlier work on functionalism [cf. note 9 (not reprinted here—*ed.*)].

   On the other hand, Putnam's articulation of social and environmental aspects of the meaning of natural-kind terms complements and supplements our viewpoint. For me, it has been a rich rewarder of reflection. More recent work of his seems to involve shifts in his viewpoint on psychological states. It may have somewhat more in common with our approach than the earlier work, but there is much that I do not understand about it.

   The argument regarding the notion of water that I extracted from Putnam's paper is narrower in scope than our argument. The Putnam-derived argument seems to work only for natural-kind terms and close relatives. And it may seem not to provide as direct a threat to certain versions of functionalism that I discuss in Section IV: At least a few philosophers would claim that one could accommodate the Putnamian argument in terms of *non*intentional formulations of input-output relations (formulations that make reference to the specific nature of the physical environment). Our argument does not submit to this maneuver. In our thought experiment, the physical environment (sofas, arthritis, and so forth in our examples) and the subject's causal relations with it (at least as these are usually conceived) were held constant. The Putnamian argument, however, has fascinatingly different implications from our argument. I have not developed these comparisons and contrasts here because doing justice to Putnam's viewpoint would demand a distracting amount of space, as the ample girth of this footnote may suggest.
3. I have discussed *de re* mental phenomena in "Belief *De Re*," *Journal of Philosophy* 74 (1977):

pp. 338–62. There I argue that all attitudes with content presuppose *de re* attitudes. Our discussion here may be seen as bearing on the details of this presupposition. But for reasons I merely sketch in the next paragraph, I think it would be a superficial viewpoint that tried to utilize our present argument to support the view that nearly all intentional mental phenomena are covertly indexical or *de re*.

The Extended Mind

Andy Clark and David J. Chalmers[1]

1. Introduction

Where does the mind stop and the rest of the world begin? The question invites two standard replies. Some accept the demarcations of skin and skull, and say that what is outside the body is outside the mind. Others are impressed by arguments suggesting that the meaning of our words 'just ain't in the head,' and hold that this externalism about meaning carries over into an externalism about mind. We propose to pursue a third position. We advocate a very different sort of externalism: an *active externalism*, based on the active role of the environment in driving cognitive processes.

2. Extended Cognition

Consider three cases of human problem-solving:

1. A person sits in front of a computer screen which displays images of various two-dimensional geometric shapes and is asked to answer questions concerning the potential fit of such shapes into depicted 'sockets.' To assess fit, the person must mentally rotate the shapes to align them with the sockets.
2. A person sits in front of a similar computer screen, but this time can choose either to physically rotate the image on the screen, by pressing a rotate button, or to mentally rotate the image as before. We can also suppose, not unrealistically, that some speed advantage accrues to the physical rotation operation.
3. Sometime in the cyberpunk future, a person sits in front of a similar computer screen. This agent, however, has the benefit of a neural implant which can perform the rotation operation as fast as the computer in the previous example. The agent must still choose which internal resource to use (the implant or the good old-fashioned mental rotation), as each resource makes different demands on attention and other concurrent brain activity.

How much *cognition* is present in these cases? We suggest that all three cases are similar. Case (3) with the neural implant seems clearly to be on a par with case (1). And case (2) with the rotation button displays the same sort of computational structure as case (3), although it is distributed across agent and computer instead of internalized within the agent. If the rotation in case (3) is cognitive, by what right do we count case (2) as fundamentally different? We cannot simply point to the skin/skull boundary as justification, since the legitimacy of that boundary is precisely what is at issue. But nothing else seems different.

The kind of case just described is by no means as exotic as it may at first appear. It is not just the presence of advanced external computing resources which raises the issue, but rather the general tendency of human reasoners to lean heavily on environmental supports. Thus consider the use of pen and paper to perform long multiplication (McClelland et al. 1986, Clark 1989), the use of physical re-arrangements of letter tiles to prompt word recall in Scrabble (Kirsh 1995), the use of instruments such as the nautical slide rule (Hutchins 1995), and the general paraphernalia of language, books, diagrams, and culture. In all these cases the

From *Analysis* 58 (1998): pp. 10–23. Reprinted with permission of the authors.

individual brain performs some operations, while others are delegated to manipulations of external media. Had our brains been different, this distribution of tasks would doubtless have varied.

In fact, even the mental rotation cases described in scenarios (1) and (2) are real. The cases reflect options available to players of the computer game Tetris. In Tetris, falling geometric shapes must be rapidly directed into an appropriate slot in an emerging structure. A rotation button can be used. David Kirsh and Paul Maglio 1994 calculate that the physical rotation of a shape through 90 degrees takes about 100 milliseconds, plus about 200 milliseconds to select the button. To achieve the same result by mental rotation takes about 1,000 milliseconds. Kirsh and Maglio go on to present compelling evidence that physical rotation is used not just to position a shape ready to fit a slot, but often to help *determine* whether the shape and the slot are compatible. The latter use constitutes a case of what Kirsh and Maglio call an 'epistemic action.' *Epistemic* actions alter the world so as to aid and augment cognitive processes such as recognition and search. Merely *pragmatic* actions, by contrast, alter the world because some physical change is desirable for its own sake (e.g., putting cement into a hole in a dam).

Epistemic action, we suggest, demands spread of *epistemic credit.* If, as we confront some task, a part of the world functions as a process which, *were it done in the head,* we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process. Cognitive processes ain't (all) in the head!

3. Active Externalism

In these cases, the human organism is linked with an external entity in a two-way interaction, creating a *coupled system* that can be seen as a cognitive system in its own right. All the components in the system play an active causal role, and they jointly govern behavior in the same sort of way that cognition usually does. If we remove the external component the system's behavioral competence will drop, just as it would if we removed part of its brain. Our thesis is that this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head.

This externalism differs greatly from standard variety advocated by Putnam 1975 and Burge 1979. When I believe that water is wet and my twin believes that twin water is wet, the external features responsible for the difference in our beliefs are distal and historical, at the other end of a lengthy causal chain. Features of the *present* are not relevant: if I happen to be surrounded by XYZ right now (maybe I have teleported to Twin Earth), my beliefs still concern standard water, because of my history. In these cases, the relevant external features are *passive.* Because of their distal nature, they play no role in driving the cognitive process in the here-and-now. This is reflected by the fact that the actions performed by me and my twin are physically indistinguishable, despite our external differences.

In the cases we describe, by contrast, the relevant external features are *active*, playing a crucial role in the here-and-now. Because they are coupled with the human organism, they have a direct impact on the organism and on its behavior. In these cases, the relevant parts of the world are *in the loop*, not dangling at the other end of a long causal chain. Concentrating on this sort of coupling leads us to an *active externalism*, as opposed to the passive externalism of Putnam and Burge.

Many have complained that even if Putnam and Burge are right about the externality of content, it is not clear that these external aspects play a causal or explanatory role in the generation of action. In counterfactual cases where internal structure is held constant but these external features are changed, behavior looks just the same; so internal structure seems to be doing the crucial work. We will not adjudicate that issue here, but we note that active externalism is not threatened by any such problem. The external features in a coupled system play an ineliminable role—if we retain internal structure but change the external features, behavior may change completely. The external features here are just as causally relevant as typical internal features of the brain.[2]

By embracing an active externalism, we allow a more natural explanation of all sorts of actions. One can explain my choice of words in Scrabble, for example, as the outcome of an extended cognitive process involving the rearrangement of tiles on my tray. Of course, one could always try to explain my action in terms of internal processes and a long series of 'inputs' and 'actions,' but this explanation would be needlessly complex. If an isomorphic process were going on in the head, we would feel no urge to characterize it in this

cumbersome way.[3] In a very real sense, the rearrangement of tiles on the tray is not part of action; it is part of *thought.*

The view we advocate here is reflected by a growing body of research in cognitive science. In areas as diverse as the theory of situated cognition (Suchman 1987), studies of real-world-robotics (Beer 1989), dynamical approaches to child development (Thelen and Smith 1994), and research on the cognitive properties of collectives of agents (Hutchins 1995), cognition is often taken to be continuous with processes in the environment.[4] Thus, in seeing cognition as extended one is not merely making a terminological decision; it makes a significant difference to the methodology of scientific investigation. In effect, explanatory methods that might once have been thought appropriate only for the analysis of 'inner' processes are now being adapted for the study of the outer, and there is promise that our understanding of cognition will become richer for it.

Some find this sort of externalism unpalatable. One reason may be that many identify the cognitive with the conscious, and it seems far from plausible that consciousness extends outside the head in these cases. But not every cognitive process, at least on standard usage, is a conscious process. It is widely accepted that all sorts of processes beyond the borders of consciousness play a crucial role in cognitive processing: in the retrieval of memories, linguistic processes, and skill acquisition, for example. So the mere fact that external processes are external where consciousness is internal is no reason to deny that those processes are cognitive.

More interestingly, one might argue that what keeps real cognition processes in the head is the requirement that cognitive processes be *portable.* Here, we are moved by a vision of what might be called the Naked Mind: a package of resources and operations we can always bring to bear on a cognitive task, regardless of the local environment. On this view, the trouble with coupled systems is that they are too easily *decoupled.* The true cognitive processes are those that lie at the constant core of the system; anything else is an add-on extra.

There is something to this objection. The brain (or brain and body) comprises a package of basic, portable, cognitive resources that is of interest in its own right. These resources may incorporate bodily actions into cognitive processes, as when we use our fingers as working memory in a tricky calculation, but they will not encompass the more contingent aspects of our external environment, such as a pocket calculator. Still, mere contingency of coupling does not rule out cognitive status. In the distant future we may be able to plug various modules into our brain to help us out: a module for extra short-term memory when we need it, for example. When a module is plugged in, the processes involving it are just as cognitive as if they had been there all along.[5]

Even if one were to make the portability criterion pivotal, active externalism would not be undermined. Counting on our fingers has already been let in the door, for example, and it is easy to push things further. Think of the old image of the engineer with a slide rule hanging from his belt wherever he goes. What if people always carried a pocket calculator, or had them implanted? The real moral of the portability intuition is that for coupled systems to be relevant to the core of cognition, *reliable* coupling is required. It happens that most reliable coupling takes place within the brain, but there can easily be reliable coupling with the environment as well. If the resources of my calculator or my Filofax are always there when I need them, then they are coupled with me as reliably as we need. In effect, they are part of the basic package of cognitive resources that I bring to bear on the everyday world. These systems cannot be impugned simply on the basis of the danger of discrete damage, loss, or malfunction, or because of any occasional decoupling: the biological brain is in similar danger, and occasionally loses capacities temporarily in episodes of sleep, intoxication, and emotion. If the relevant capacities are generally there when they are required, this is coupling enough.

Moreover, it may be that the biological brain has in fact evolved and matured in ways which factor in the reliable presence of a manipulable external environment. It certainly seems that evolution has favored on-board capacities which are especially geared to parasitizing the local environment so as to reduce memory load, and even to transform the nature of the computational problems themselves. Our visual systems have evolved to rely on their environment in various ways: they exploit contingent facts about the structure of natural scenes (e.g. Ullman and Richards 1984), for example, and they take advantage of the computational shortcuts afforded by bodily motion and locomotion (e.g. Blake and Yuille 1992). Perhaps there are other cases where evolution has found it advantageous to exploit the possibility of the environment being in the cognitive loop. If so,

then external coupling is part of the truly basic package of cognitive resources that we bring to bear on the world.

Language may be an example. Language appears to be a central means by which cognitive processes are extended into the world. Think of a group of people brainstorming around a table, or a philosopher who thinks best by writing, developing her ideas as she goes. It may be that language evolved, in part, to enable such extensions of our cognitive resources within actively coupled systems.

Within the lifetime of an organism, too, individual learning may have molded the brain in ways that rely on cognitive extensions that surrounded us as we learned. Language is again a central example here, as are the various physical and computational artifacts that are routinely used as cognitive extensions by children in schools and by trainees in numerous professions. In such cases the brain develops in a way that complements the external structures, and learns to play its role within a unified, densely coupled system. Once we recognize the crucial role of the environment in constraining the evolution and development of cognition, we see that extended cognition is a core cognitive process, not an add-on extra.

An analogy may be helpful. The extraordinary efficiency of the fish as a swimming device is partly due, it now seems, to an evolved capacity to couple its swimming behaviors to the pools of external kinetic energy found as swirls, eddies and vortices in its watery environment (see M. Triantafyllou and G. Triantafyllou 1995). These vortices include both naturally occurring ones (e.g., where water hits a rock) and self-induced ones (created by well-timed tail flaps). The fish swims by building these externally occurring processes into the very heart of its locomotion routines. The fish and surrounding vortices together constitute a unified and remarkably efficient swimming machine.

Now consider a reliable feature of the human environment, such as the sea of words. This linguistic surround envelops us from birth. Under such conditions, the plastic human brain will surely come to treat such structures as a reliable resource to be factored into the shaping of on-board cognitive routines. Where the fish flaps its tail to set up the eddies and vortices it subsequently exploits, we intervene in multiple linguistic media, creating local structures and disturbances whose reliable presence drives our ongoing internal processes. Words and external symbols are thus paramount among the cognitive vortices which help constitute human thought.

4. From Cognition to Mind

So far we have spoken largely about 'cognitive processing,' and argued for its extension into the environment. Some might think that the conclusion has been bought too cheaply. Perhaps some *processing* takes place in the environment, but what of *mind?* Everything we have said so far is compatible with the view that truly mental states—experiences, beliefs, desires, emotions, and so on—are all determined by states of the brain. Perhaps what is truly mental is internal, after all?

We propose to take things a step further. While some mental states, such as experiences, may be determined internally, there are other cases in which external factors make a significant contribution. In particular, we will argue that *beliefs* can be constituted partly by features of the environment, when those features play the right sort of role in driving cognitive processes. If so, the mind extends into the world.

First, consider a normal case of belief embedded in memory. Inga hears from a friend that there is an exhibition at the Museum of Modern Art, and decides to go see it. She thinks for a moment and recalls that the museum is on 53rd Street, so she walks to 53rd Street and goes into the museum. It seems clear that Inga believes that the museum is on 53rd Street, and that she believed this even before she consulted her memory. It was not previously an *occurrent* belief, but then neither are most of our beliefs. The belief was sitting somewhere in memory, waiting to be accessed.

Now consider Otto. Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. Today, Otto hears about the exhibition at the Museum of Modern Art, and decides to go see it. He consults the notebook, which says that the museum is on 53rd Street, so he walks to 53rd Street and goes into the museum.

Clearly, Otto walked to 53rd Street because he wanted to go to the museum and he believed the museum was on 53rd Street. And just as

Inga had her belief even before she consulted her memory, it seems reasonable to say that Otto believed the museum was on 53rd Street even before consulting his notebook. For in relevant respects the cases are entirely analogous: the notebook plays for Otto the same role that memory plays for Inga. The information in the notebook functions just like the information constituting an ordinary non-occurrent belief; it just happens that this information lies beyond the skin.

The alternative is to say that Otto has no belief about the matter until he consults his notebook; at best, he believes that the museum is located at the address in the notebook. But if we follow Otto around for a while, we will see how unnatural this way of speaking is. Otto is constantly using his notebook as a matter of course. It is central to his actions in all sorts of contexts, in the way that an ordinary memory is central in an ordinary life. The same information might come up again and again, perhaps being slightly modified on occasion, before retreating into the recesses of his artificial memory. To say that the beliefs disappear when the notebook is filed away seems to miss the big picture in just the same way as saying that Inga's beliefs disappear as soon as she is no longer conscious of them. In both cases the information is reliably there when needed, available to consciousness and available to guide action, in just the way that we expect a belief to be.

Certainly, insofar as beliefs and desires are characterized by their explanatory roles, Otto's and Inga's cases seem to be on a par: the essential causal dynamics of the two cases mirror each other precisely. We are happy to explain Inga's action in terms of her occurrent desire to go to the museum and her standing belief that the museum is on 53rd street, and we should be happy to explain Otto's action in the same way. The alternative is to explain Otto's action in terms of his occurrent desire to go to the museum, his standing belief that the Museum is on the location written in the notebook, and the accessible fact that the notebook says the Museum is on 53rd Street; but this complicates the explanation unnecessarily. If we must resort to explaining Otto's action this way, then we must also do so for the countless other actions in which his notebook is involved; in each of the explanations, there will be an extra term involving the notebook. We submit that to explain things this way is to take *one step too many*. It is pointlessly complex, in the same way that it would be pointlessly complex to explain Inga's actions in terms of beliefs about her memory. The notebook is a constant for Otto, in the same way that memory is a constant for Inga; to point to it in every belief/desire explanation would be redundant. In an explanation, simplicity is power.

If this is right, we can even construct the case of Twin Otto, who is just like Otto except that a while ago he mistakenly wrote in his notebook that the Museum of Modern Art was on 51st Street. Today, Twin Otto is a physical duplicate of Otto from the skin in, but his notebook differs. Consequently, Twin Otto is best characterized as believing that the museum is on 51st Street, where Otto believes it is on 53rd. In these cases, a belief is simply not in the head.

This mirrors the conclusion of Putnam and Burge, but again there are important differences. In the Putnam/Burge cases, the external features constituting differences in belief are distal and historical, so that twins in these cases produce physically indistinguishable behavior. In the cases we are describing, the relevant external features play an active role in the here-and-now, and have a direct impact on behavior. Where Otto walks to 53rd Street, Twin Otto walks to 51st. There is no question of explanatory irrelevance for this sort of external belief content; it is introduced precisely because of the central explanatory role that it plays. Like the Putnam and Burge cases, these cases involve differences in reference and truth-conditions, but they also involve differences in the dynamics of *cognition*.[6]

The moral is that when it comes to belief, there is nothing sacred about skull and skin. What makes some information count as a belief is the role it plays, and there is no reason why the relevant role can be played only from inside the body.

Some will resist this conclusion. An opponent might put her foot down and insist that as she uses the term 'belief,' or perhaps even according to standard usage, Otto simply does not qualify as believing that the museum is on 53rd Street. We do not intend to debate what is standard usage; our broader point is that the notion of belief *ought* to be used so that Otto qualifies as having the belief in question. In all *important* respects, Otto's case is similar to a standard case of (non-occurrent) belief. The differences between Otto's case and Inga's are striking, but they are superficial. By using the 'belief' notion in a wider way, it picks out something more akin to a natural kind. The notion becomes deeper and more unified, and is more useful in explanation.

To provide substantial resistance, an opponent has to show that Otto's and Inga's cases differ in some important and relevant respect. But in what deep respect are the cases different? To make the case *solely* on the grounds that information is in the head in one case but not in the other would be to beg the question. If this difference is relevant to a difference in belief, it is surely not *primitively* relevant. To justify the different treatment, we must find some more basic underlying difference between the two.

It might be suggested that the cases are relevantly different in that Inga has more *reliable* access to the information. After all, someone might take away Otto's notebook at any time, but Inga's memory is safer. It is not implausible that constancy is relevant: indeed, the fact that Otto always uses his notebook played some role in our justifying its cognitive status. If Otto were consulting a guidebook as a one-off, we would be much less likely to ascribe him a standing belief. But in the original case, Otto's access to the notebook is very reliable—not perfectly reliable, to be sure, but then neither is Inga's access to her memory. A surgeon might tamper with her brain, or more mundanely, she might have too much to drink. The mere possibility of such tampering is not enough to deny her the belief.

One might worry that Otto's access to his notebook *in fact* comes and goes. He showers without the notebook, for example, and he cannot read it when it is dark. Surely his belief cannot come and go so easily? We could get around this problem by redescribing the situation, but in any case an occasional temporary disconnection does not threaten our claim. After all, when Inga is asleep, or when she is intoxicated, we do not say that her belief disappears. What really counts is that the information is easily available when the subject needs it, and this constraint is satisfied equally in the two cases. If Otto's notebook were often unavailable to him at times when the information in it would be useful, there might be a problem, as the information would not be able to play the action-guiding role that is central to belief; but if it is easily available in most relevant situations, the belief is not endangered.

Perhaps a difference is that Inga has *better* access to the information than Otto does? Inga's 'central' processes and her memory probably have a relatively high-bandwidth link between them, compared to the low-grade connection between Otto and his notebook. But this alone does not make a difference between believing and not believing. Consider Inga's museum-going friend Lucy, whose biological memory has only a low-grade link to her central systems, due to nonstandard biology or past misadventures. Processing in Lucy's case might be less efficient, but as long as the relevant information is accessible, Lucy clearly believes that the museum is on 53rd Street. If the connection was too indirect—if Lucy had to struggle hard to retrieve the information with mixed results, or a psychotherapist's aid were needed—we might become more reluctant to ascribe the belief, but such cases are well beyond Otto's situation, in which the information is easily accessible.

Another suggestion could be that Otto has access to the relevant information only by *perception*, whereas Inga has more direct access—by introspection, perhaps. In some ways, however, to put things this way is to beg the question. After all, we are in effect advocating a point of view on which Otto's internal processes and his notebook constitute a single cognitive system. From the standpoint of this system, the flow of information between notebook and brain is not perceptual at all; it does not involve the impact of something outside the system. It is more akin to information flow within the brain. The only deep way in which the access is perceptual is that in Otto's case, there is a distinctly perceptual phenomenology associated with the retrieval of the information, whereas in Inga's case there is not. But why should the nature of an associated phenomenology make a difference to the status of a belief? Inga's memory may have some associated phenomenology, but it is still a belief. The phenomenology is not visual, to be sure. But for visual phenomenology consider the Terminator, from the Arnold Schwarzenegger movie of the same name. When he recalls some information from memory, it is 'displayed' before him in his visual field (presumably he is conscious of it, as there are frequent shots depicting his point of view). The fact that standing memories are recalled in this unusual way surely makes little difference to their status as standing beliefs.

These various small differences between Otto's and Inga's cases are all *shallow* differences. To focus on them would be to miss the way in which for Otto, notebook entries play just the sort of role that beliefs play in guiding most people's lives.

Perhaps the intuition that Otto's is not a true belief comes from a residual feeling that the only true beliefs are occurrent beliefs. If we take this feeling seriously, Inga's belief

will be ruled out too, as will many beliefs that we attribute in everyday life. This would be an extreme view, but it may be the most consistent way to deny Otto's belief. Upon even a slightly less extreme view—the view that a belief must be *available* for consciousness, for example—Otto's notebook entry seems to qualify just as well as Inga's memory. Once dispositional beliefs are let in the door, it is difficult to resist the conclusion that Otto's notebook has all the relevant dispositions.

5. Beyond the Outer Limits

If the thesis is accepted, how far should we go? All sorts of puzzle cases spring to mind. What of the amnesic villagers in *100 Years of Solitude*, who forget the names for everything and so hang labels everywhere? Does the information in my Filofax count as part of my memory? If Otto's notebook has been tampered with, does he believe the newly-installed information? Do I believe the contents of the page in front of me before I read it? Is my cognitive state somehow spread across the Internet?

We do not think that there are categorical answers to all of these questions, and we will not give them. But to help understand what is involved in ascriptions of extended belief, we can at least examine the features of our central case that make the notion so clearly applicable there. First, the notebook is a constant in Otto's life—in cases where the information in the notebook would be relevant, he will rarely take action without consulting it. Second, the information in the notebook is directly available without difficulty. Third, upon retrieving information from the notebook he automatically endorses it. Fourth, the information in the notebook has been consciously endorsed at some point in the past, and indeed is there as a consequence of this endorsement.[7] The status of the fourth feature as a criterion for belief is arguable (perhaps one can acquire beliefs through subliminal perception, or through memory tampering?), but the first three features certainly play a crucial role.

Insofar as increasingly exotic puzzle cases lack these features, the applicability of the notion of 'belief' gradually falls off. If I rarely take relevant action without consulting my Filofax, for example, its status within my cognitive system will resemble that of the notebook in Otto's. But if I often act without consultation—for example, if I sometimes answer relevant questions with 'I don't know'—then information in it counts less clearly as part of my belief system. The Internet is likely to fail on multiple counts, unless I am unusually computer-reliant, facile with the technology, and trusting, but information in certain files on my computer may qualify. In intermediate cases, the question of whether a belief is present may be indeterminate, or the answer may depend on the varying standards that are at play in various contexts in which the question might be asked. But any indeterminacy here does not mean that in the central cases, the answer is not clear.

What about socially extended cognition? Could my mental states be partly constituted by the states of other thinkers? We see no reason why not, in principle. In an unusually interdependent couple, it is entirely possible that one partner's beliefs will play the same sort of role for the other as the notebook plays for Otto.[8] What is central is a high degree of trust, reliance, and accessibility. In other social relationships these criteria may not be so clearly fulfilled, but they might nevertheless be fulfilled in specific domains. For example, the waiter at my favorite restaurant might act as a repository of my beliefs about my favorite meals (this might even be construed as a case of extended desire). In other cases, one's beliefs might be embodied in one's secretary, one's accountant, or one's collaborator.[9]

In each of these cases, the major burden of the coupling between agents is carried by language. Without language, we might be much more akin to discrete Cartesian 'inner' minds, in which high-level cognition relies largely on internal resources. But the advent of language has allowed us to spread this burden into the world. Language, thus construed, is not a mirror of our inner states but a complement to them. It serves as a tool whose role is to extend cognition in ways that on-board devices cannot. Indeed, it may be that the intellectual explosion in recent evolutionary time is due as much to this linguistically-enabled extension of cognition as to any independent development in our inner cognitive resources.

What, finally, of the self? Does the extended mind imply an extended self? It seems so. Most of us already accept that the self outstrips the boundaries of consciousness; my dispositional beliefs, for example, constitute in some deep sense part of who I am. If so, then these boundaries may also fall beyond the skin. The information in Otto's notebook, for example, is a central part of his identity as a cognitive agent. What this comes to is that Otto *himself* is best

regarded as an extended system, a coupling of biological organism and external resources. To consistently resist this conclusion, we would have to shrink the self into a mere bundle of occurrent states, severely threatening its deep psychological continuity. Far better to take the broader view, and see agents themselves as spread into the world.

As with any reconception of ourselves, this view will have significant consequences. There are obvious consequences for philosophical views of the mind and for the methodology of research in cognitive science, but there will also be effects in the moral and social domains. It may be, for example, that in some cases interfering with someone's environment will have the same moral significance as interfering with their person. And if the view is taken seriously, certain forms of social activity might be reconceived as less akin to communication and action, and as more akin to thought. In any case, once the hegemony of skin and skull is usurped, we may be able to see ourselves more truly as creatures of the world.

REFERENCES

Beer, R. *Intelligence as Adaptive Behavior* (New York: Academic Press, 1989).

Blake, A., and Yuille, A., eds., *Active Vision* (Cambridge, MA: MIT Press. 1992).

Burge, T., "Individualism and the mental," *Midwest Studies in Philosophy* 4 (1979): pp. 73–122.

Clark, A. *Microcognition* (Cambridge, MA: MIT Press, 1989).

Haugeland, J., "Mind embodied and embedded," in *Mind and Cognition*, Y. Houng and J. Ho, eds., (Taipei: Academia Sinica, 1995).

Hutchins, E. *Cognition in the Wild* (Cambridge, MA: MIT Press, 1995).

Kirsh, D., "The intelligent use of space," *Artificial Intelligence* 73 (1995): pp. 31–68.

Kirsh, D., and Maglio, P., "On distinguishing epistemic from pragmatic action," *Cognitive Science* 18 (1994): pp. 513–49.

McClamrock, R. *Existential Cognition* (Chicago: University of Chicago Press, 1995).

McClelland, J. L., D. E. Rumelhart, and G. E. Hinton, "The appeal of parallel distributed processing," in *Parallel Distributed Processing*, vol. 2, J. L. McClelland and D. E. Rumelhart, eds., (Cambridge, MA: MIT Press, 1986).

McHugh, M. *China Mountain Zhang* (New York: Tom Doherty Associates, 1992).

Putnam, H., "The meaning of 'meaning'," in *Language, Mind, and Knowledge*, K. Gunderson, ed., (Minneapolis: University of Minnesota Press, 1975).

Simon, H. *The Sciences of the Artificial* (Cambridge, MA: MIT Press, 1981).

Suchman, L. *Plans and Situated Actions* (Cambridge, UK: Cambridge University Press, 1987).

Thelen, E., and Smith, L. *A Dynamic Systems Approach to the Development of Cognition and Action* (Cambridge, MA: MIT Press, 1994).

Triantafyllou, M., and Triantafyllou, G., "An Efficient Swimming Machine," *Scientific American* 272, no. 3 (1995): pp. 64–70.

Ullman, S., and Richards, W. *Image Understanding* (Norwood, NJ: Ablex, 1984).

Varela, F., Thompson, E., and Rosch, E. *The Embodied Mind* (Cambridge, MA: MIT Press. 1991).

Wilson, R., "Wide computationalism," *Mind* 103 (1994): no. 351–72.

NOTES

1. Authors are listed in order of degree of belief in the central thesis.
2. Much of the appeal of externalism in the philosophy of mind may stem from the intuitive appeal of active externalism. Externalists often make analogies involving external features in coupled systems, and appeal to the arbitrariness of boundaries between brain and environment. But these intuitions sit uneasily with the letter of standard externalism. In most of the Putnam/Burge cases, the immediate environment is irrelevant; only the historical environment counts. Debate has focused on the question of whether mind must be in the head, but a more relevant question in assessing these examples might be: is mind in the present?
3. Herbert Simon 1981 once suggested that we view internal memory as, in effect, an external resource upon which 'real' inner processes operate. 'Search in memory,' he comments, 'is not very different from search of the external environment.' Simon's view at least has the virtue of treating internal and external processing with the parity they deserve, but we suspect that on his view the mind will shrink too small for most people's tastes.
4. Philosophical views of a similar spirit can be found in Haugeland 1995, McClamrock 1985, Varela et al. 1991, and Wilson 1994.
5. Or consider the following passage from a recent science fiction novel (McHugh 1992, p. 213): 'I am taken to the system's department where I am attuned to the system. All I do is jack in and then a technician instructs the system to attune and it does. I jack out and query the time. 10:52. The information pops up. Always before I could only access information when I was jacked in, it gave me a sense that I knew what I thought and what the system told

me, but now, how do I know what is system and what is Zhang?'

6. In the terminology of Chalmers' "The Components of Content": the twins in the Putnam and Burge cases differ only in their *subjunctive* content (or relational content), but Otto and his twin can be seen to differ in their *epistemic* content (or notional content), which is the sort of content that governs cognition. Epistemic content is generally internal to a cognitive system, but in this case the cognitive system is itself effectively extended to include the notebook.
7. The constancy and past-endorsement criteria may suggest that history is partly constitutive of belief. One might react to this by removing any historical component (giving a purely dispositional reading of the constancy criterion and eliminating the past-endorsement criterion, for example), or one might allow such a component as long as the main burden is carried by features of the present.
8. From the *New York Times*, March 30, 1995, p. B7, in an article on former UCLA basketball coach John Wooden: 'Wooden and his wife attended 36 straight Final Fours, and she invariably served as his memory bank. Nell Wooden rarely forgot a name—her husband rarely remembered one—and in the standing-room-only Final Four lobbies, she would recognize people for him.'
9. Might this sort of reasoning also allow something like Burge's extended 'arthritis' beliefs? After all, I might always defer to my doctor in taking relevant actions concerning my disease. Perhaps so, but there are some clear differences. For example, any extended beliefs would be grounded in an existing active relationship with the doctor, rather than in a historical relationship to a language community. And on the current analysis, my deference to the doctor would tend to yield something like a true belief that I have some other disease in my thigh, rather than the false belief that I have arthritis there. On the other hand, if I used medical experts solely as terminological consultants, the results of Burge's analysis might be mirrored.

Overextending the Mind

Brie Gertler

Clark and Chalmers argue that the mind is *extended*—that is, its boundary lies beyond the skin. (Clark and Chalmers 1998, reprinted as chapter 49 of this volume.[1] For brevity, I will refer to the authors as 'C&C'). In this essay, I will criticize this conclusion. However, I will also defend some of the more controversial elements of C&C's argument. I reject their conclusion because I think that their argument shows that a seemingly innocuous assumption, about *internal* states and processes, is flawed.

The first section of the essay outlines C&C's argument. In Section 2, I sketch some unpalatable consequences of their conclusion. Insofar as we want to avoid these consequences, we should look for a flaw in the argument. As outlined in Section 1, the argument appears to be valid, so finding a flaw means identifying a premise that it is reasonable to reject. In Section 3, I evaluate each of the major premises of the argument and find that all but one are acceptable; I then explain why I reject the remaining premise. Section 4 briefly defends the picture of the mind that emerges from rejecting this premise.

My goal is not to conclusively refute C&C's argument. My aim is only to reveal the best alternative for those who remain skeptical about the existence—or, perhaps, even the possibility—of extended minds.

1. Clark and Chalmers' argument

The authors provide two arguments to show that the mind is extended. First, they argue that the mind's cognitive processes can at least partially consist in processes performed by external devices. Their examples of such external cognitive processing devices include a computer that you can use to rotate shapes when playing the game Tetris. As they describe this case, the computer's rotation of a shape plays the same sort of role, in your cognitive economy, as the corresponding internal process (when you simply

From B. Gertler and L. Shapiro (eds.) *Arguing About the Mind* (Routledge, 2007).

imagine how the shape would appear if it were rotated in various ways). For instance, the result of this process is automatically endorsed—you believe that the shape would look like *that* when rotated. And you use this information to guide your behavior, such as moving the joystick to position the shape in a certain place on the screen. They conclude that insofar as the internal process of imagining qualifies as *your cognitive process*, so should the external computational process.

While I will return to this processing case at various points below, my remarks will focus on the second of C&C's arguments: that standing beliefs (and desires, etc.) can be partially constituted by factors external to the skin. Standing beliefs include stored memories and other beliefs that are not currently being entertained. The notion of a *standing* belief contrasts with the notion of an *occurrent* belief, which is a conviction that you are now entertaining. For instance, you probably have the standing belief that dinosaurs once roamed the earth. At the moment before you read that sentence, the belief was *simply* a standing belief; it was not occurrent (unless you happened to be thinking about dinosaurs at that moment). But now that you're thinking about the fact that dinosaurs roamed the earth, that belief is *occurrent*.

C&C's principal examples of extended standing beliefs involve a character they call Otto. Otto, who suffers from Alzheimer's disease, carries a notebook in which he routinely records useful information of the sort that most of us would easily commit to memory. Otto consults the notebook whenever he needs this stored information to guide his reasoning or actions. For instance, on a trip to the Museum of Modern Art in New York, Otto frequently consults the notebook, to remind himself that he is going to the MoMA, that the MoMA is on 53rd Street, etc. C&C claim that the information stored in Otto's notebook—such as 'the MoMA is on 53rd Street'—partially constitutes his standing beliefs, and hence that his mind extends beyond his skin.

Here is my reconstruction of C&C's argument.

(1) 'What makes some information count as a [standing] belief is the role it plays' [p. xx14xx].
(2) 'The information in the notebook functions just like [that is, it plays the same role as] the information constituting an ordinary non-occurrent belief' [p. xx13xx].
(3) The information in Otto's notebook counts as standing beliefs[2] (from [1] and [2]).
(4) Otto's standing beliefs are part of his mind.
(5) The information in Otto's notebook is part of Otto's mind (from [3] and [4]).
(6) Otto's notebook belongs to the world external to Otto's skin, i.e., the 'external' world.
(7) The mind extends into the world (from [5] and [6]).

In assessing C&C's extended mind hypothesis, I will focus on the conclusion that Otto's standing beliefs extend into the world. Later, I will briefly discuss how my assessment applies to the case of cognitive processing.

2. Some worrisome consequences of Clark and Chalmers' conclusion

C&C's conclusion is that 'the mind extends into the world' where 'the world' refers to what is beyond the subject's skin. In this section, I will use the example of Otto and his notebook to describe two consequences that seem to follow from this conclusion. Both of these consequences are, I think, worrisome; the second is especially so. Recognizing them will thus cast doubt on the conclusion.

First consequence: limits on introspection

It is commonly held that, in general, a subject can determine his or her own beliefs and desires by using a method that others cannot use (to determine that subject's beliefs). Let us use the term 'introspection' to refer to this method. Introspection is, in this sense, a *necessarily* first-person method: it reveals only the introspector's own states, and not the states of others. Introspection may not be infallible; in fact, it may be no more reliable than third-person methods. The claim is only that each of us has a *way* of gaining access to our own beliefs that is unavailable to others.

According to C&C, the information in Otto's notebook partially constitutes some of his standing beliefs. Can Otto introspect these beliefs, in our sense of 'introspect'? That is, can he identify these beliefs by using a method available only to himself?

I think that he cannot. When Otto tries to figure out what he believes on a particular

topic, he consults the notebook. For instance, suppose that he wonders what he believes about the location of the MoMA. He will look in the notebook and conclude: *I believe that the MoMA is on 53rd Street.* But of course someone other than Otto can determine Otto's beliefs in precisely the same way: by consulting the notebook, a friend can determine that Otto believes that the MoMA is on 53rd Street. So it appears that, if the entries in Otto's notebook partially constitute his beliefs, then Otto cannot introspect his beliefs.

Much more could be said here. For one thing, it might be argued that when Otto consults the notebook, in order to determine what he believes about the location of the museum, he *is* introspecting. C&C seem to suggest this when they say that treating Otto's access to the notebook as perceptual rather than introspective would beg the question against the claim that the notebook is *part of* Otto's mind (p. xx16xx). But as I am using this term, 'introspection' refers only to those processes that are *necessarily* first-personal. Someone who claimed that, in consulting the notebook, Otto is introspecting in *my* sense would have to show that Otto has a unique kind of access to the notebook—or, perhaps, to the fact that the notebook entries play the relevant 'belief' role in his cognitive economy. But it is difficult to see how this access could be unique, so long as it was access to a feature *external* to Otto's skin.

Another possibility that C&C describe more directly reveals the lack of unique first-person access.

> In an unusually interdependent couple, it is entirely possible that one partner's beliefs will play the same sort of role for the other as the notebook plays for Otto. . . . [O]ne's beliefs might be embodied in one's secretary, one's accountant, or one's collaborator [pp. xx17–18xx].

To flesh out this scenario, suppose that Amanda, an absent-minded executive, uses her assistant Fred as a repository of her daily schedule. Fred knows that Amanda has a 2:00 board meeting on Monday, and stores this information for Amanda. Since this information plays the appropriate role in Amanda's cognitive economy (it is readily accessible to her, automatically endorsed by her, etc.), it counts as her belief.

Now suppose that Amanda wonders what she believes about her Monday schedule. To determine this, she will consult Fred, to see what he believes about it. But this is the same process that Fred uses to determine what Amanda believes about her Monday schedule. Recognizing that he is a repository for Amanda's standing beliefs, Fred will determine Amanda's beliefs about the schedule simply by consulting his own beliefs about it. Amanda's access to her beliefs, and to the fact that she has those beliefs, proceeds via a method also available to Fred. So Amanda has no uniquely first-personal method of determining what she believes; that is, she cannot introspect her beliefs, in my sense of 'introspect'

C&C would likely accept this consequence. They could simply allow that, in general, we have unique introspective access only to our *occurrent* experiences and our *occurrent* thoughts, that is, thoughts that we are now entertaining. (Crucially, they do not claim that occurrent thoughts are extended.) The point may be even clearer when applied to cognitive processes such as those involved in the Tetris case. You do not seem to have any special first-person access to *how* you go about imagining the shape rotated: you simply perform this feat of imagination.[3] So C&C can easily allow that those states that are extended—such as standing beliefs and nonconscious cognitive processes—are simply non-introspectible.

Still, for one who thinks that *introspectibility* is crucial to our basic concept of the mind, this point will cast doubt on C&C's conclusion that the mind extends into the world. If one can introspect only the non-extended parts of the mind, then why count the external factors as truly part of the mind? (I will return to this point in Section 4.)

I now turn to the second, more troubling consequence of C&C's conclusion.

Second consequence: a proliferation of actions

C&C dub their view 'active externalism' to highlight what they see as one of its chief benefits: the extended states that it counts as *mental* play a crucial role in generating action. In marking this benefit, they appear to suggest that this contribution to action is, at least in part, what justifies counting the wide states as truly *mental* states. But it's not clear that the wide states play the crucial role C&C ascribe to them.[4]

A simple thought experiment will convey the basis for doubt on this point. Suppose that, instead of a notebook, Otto uses an external computing device as a repository for important information. Suppose also that he records

some of his desires in the device. For instance, he records the desire to make banana bread on Tuesday; the belief that banana bread requires bananas; the belief that the corner grocery store is a good source for bananas; etc. And he allows the device to perform some cognitive processes for him, including devising action plans based on the information it has stored. (C&C would surely allow that a single device could both serve as a repository for standing states and perform cognitive processes as in the Tetris example; after all, the brain accomplishes both of these tasks.) The idea that external devices can devise action plans is nothing new. For example, a dashboard-mounted Global Positioning System records the subject's desire to reach a particular destination, and uses stored geographical information to devise the most efficient route to fulfilling that desire.

Finally, imagine that this computing device is plugged into a humanoid robot that Otto also owns. In effect, the computing device serves as part of the robot 'brain' (Otto's internal, *organic* brain may be another part of the robot's brain.) It uses inputs from the robot's various detection systems to determine the layout of its environment, and it controls the robot's movements by sending signals to the robot's 'limbs.'

Otto spends Monday asleep in bed. (Or, rather, the organic portion of his body does—after all, if C&C are correct the external device qualifies as part of Otto's body.) The robot is, however, very active: using the information stored within it, it 'realizes' that a trip to the grocery store is in order, since this is the most efficient way to execute the desire to make banana bread on Tuesday. Drawing on various other bits of information, it goes to the grocery store, purchases bananas, and returns home. Alas, the organism's sleep is very deep, and he (it?) does not awake until late on Tuesday. When he does, he is roused by the tantalizing scent of freshly baked banana bread.

Now did *Otto* make the bread? It seems that C&C should say that he did. They claim that, in explaining why Otto walked to 53rd Street, we need not cite the occurrent belief that the MoMA is on 53rd Street, which Otto has (for a fleeting moment) upon consulting the notebook. Instead, they say, an adequate explanation may simply cite the notebook entry itself. Expanding on this claim, it seems that in order to explain the bread-making behavior, we need not cite any occurrent belief or desire of Otto's; we can simply cite the information and dispositions stored in the robot's 'brain.' The implication of premises (1) and (2) of their argument is that these 'count as' Otto's standing beliefs and desires. So long as no occurrent belief or desire needs to be cited in an action explanation, the bits of behavior that directly result from the bits of information stored in the robot—the trip to the grocery store, the making of the banana bread—seem accurately described as *Otto's actions*.

To resist this, it may be argued that Otto *himself* didn't go to the grocery store, or make the banana bread—for he was asleep in bed. But notice that this reply depends on denying that the robot is part of Otto. And if C&C are correct, there is little support for such a distinction. 'Otto *himself* is best regarded as an extended system, a coupling of biological organism and external resources' (p. xx18xx). Surely, some of Otto's actions directly involve only the organic part of Otto; so, on grounds of parity, we should allow that some of Otto's actions might directly involve only his non-organic part.

Of course, the organism and the robot will constitute the right kind of extended system only if they are related in certain ways. But they do seem to be appropriately related. The information stored in the robot's 'brain' is present because the organism occurrently believed it in the past; it is readily accessible; the organism automatically endorses the information it contains; etc.[5] These are the conditions that are met by Otto's notebook and that, according to C&C, make it the case that Otto's mind extends to the notebook. On the same grounds, then, we should say that in our story, Otto extends to the external device. We can even imagine that when the organism awakens Otto compliments himself on his baking prowess.

So C&C's argument suggests that making the banana bread was Otto's action. If this is correct, there seems no limit to the actions a single person can perform. For imagine that organic Otto programs an enormous fleet of robots, linking them so that they are in constant communication with each other. These robots then engage in widespread, multifarious activities. Some take a slow boat to China; others descend on a neighborhood in Texas and ring all the doorbells at once; others compete in karaoke contests in Tokyo. When we say that all of these activities are Otto's actions, we are not simply saying that he is *somehow* responsible for them, or that he did something in the past that causally contributed to them. We are saying that he is, quite literally, *performing* each of these actions: he is enormously busy (though tireless)

and, unlike Superman, he can be in two places at once. Given that Otto's standing states might extend to a notebook, they can also extend to indefinitely many other external devices. Add to this the claim that Otto's actions may be the product of his standing states alone, and Otto becomes extraordinarily active.

If this result seems implausible, then perhaps we should question whether behavior that results from these extended states truly qualifies as action. We might limit *actions* in various ways: e.g., by requiring that the organic part of one's body must be involved in a genuine action, or by requiring that an agent's occurrent beliefs and desires (which, they assume, are internal to the organic body) be involved in each of her genuine actions. But these moves conflict with C&C's claim that there is nothing *special*, vis-à-vis a subject's agency, about states internal to her organic body.

This second consequence, that actions occur at a distance from the organic agent and proliferate excessively, is more threatening to C&C's conclusion than the first consequence. I observed that C&C could simply accept that extended states cannot be introspected, while maintaining that they are nevertheless mental. On that view, the distinction between what is introspectible and what is not may parallel the distinction between what is internal and what is extended. Since they claim that some non-introspectible (extended) states and processes are mental, they must use some other feature to distinguish the mental from the non-mental. And given that they stress the 'active' nature of these extended states and processes, the relevant feature seems to be this: extended states and processes qualify as mental because a piece of behavior caused by such states and processes (even one in which no occurrent states or processes play a crucial role) qualifies as a genuine, intentional *action*. So if we are hesitant to describe making the banana bread as *Otto's action*, then we have serious grounds for doubt that extended states and processes have the feature that, on C&C's view, qualifies them as truly *mental*.

Both of these worrisome consequences derive from an exceedingly liberal conception of mind. If we can restrict the mind—to that which is introspectible, or to states that causally explain bits of behavior that (unlike the robot's behavior) seem like genuine *actions*—then we can avoid both of these consequences.

But where did C&C go wrong? Which step of their argument is responsible for this problematic inflation of the mental?

3. Where is the flaw in C&C's argument?

If one or both of the consequences just outlined seem objectionable, we should find a way to block C&C's argument for the extended mind hypothesis. As it is outlined in Section 1, their argument seems valid. So we can block its conclusion only by finding fault with at least one of its premises. Two of the premises—namely, (3) and (5)—follow from other premises. To find the flaw in the argument, then, we must look to premises (1), (2), (4), and (6).

We can quickly rule out premise (6) as the source of the problem. For this premise merely stipulates that 'the world' in the conclusion, refers to the part of reality that is *beyond the skin*. We can have little quarrel with such a stipulation.

The major premises remaining are (1), (2), and (4). Let us examine each of these in turn.

Here is premise (1):

> (1) 'What makes some information count as a [standing] belief is the role it plays'

Premise (1) suggests that standing beliefs are defined, as such, by their role—loosely speaking, by how they *function*. It thus amounts to a kind of *functionalism* about standing beliefs. Given premise (1)'s contribution to this argument, the best way to reject it would be to show that, while playing a certain functional role may be *necessary* for being a standing belief, it is not *sufficient*: standing beliefs must not only play the relevant functional role, but must possess certain other features as well. This would mean that, even if the information in Otto's notebook played the same functional role as his standing beliefs, it might not constitute his standing beliefs because it lacked those other features. Obvious possibilities for such additional features include being constituted by a certain kind of material (e.g., the organic grey matter that makes up the human brain), or being located entirely within the skin. If we require that standing beliefs have one or both of these features, then we will reject premise (1).

C&C resist these further requirements, and many other philosophers would agree with them on this point. Limiting standing beliefs to states that are constituted by a particular kind of *material* seems unacceptably ad hoc: for instance, it would exclude the possibility that extraterrestrials with very different physical constitutions could have standing beliefs. And the claim that

standing beliefs must be located within the skin is question-begging, for the significance of the 'skin' boundary is precisely what is at issue here. So we cannot assume that standing beliefs must be internal.

Of course, rather than simply assuming that standing beliefs must be internal, one might *argue* for this claim. The most straightforward way to do this is to reject premise (2), viz.:

(2) 'The information in the notebook functions just like [that is, it plays the same role as] the information constituting an ordinary non-occurrent belief'

If the information in the notebook does not play the same role as ordinary internal beliefs, this may help to show that the particular 'belief' role can be played only by internal states. So let us turn to examining the prospects for premise (2).

C&C outline the functional role that both the notebook entries and ordinary (internal) standing beliefs play, with the following four conditions.[6]

(i) They are consistently available.
(ii) They are readily accessible.
(iii) They are automatically endorsed.
(iv) They are present (in the brain or notebook) because they were consciously endorsed in the past.

Now according to premise (2), the notebook entries play the same role as ordinary internal standing beliefs. There are two ways to reject this premise. We can deny that the notebook entries meet (i)–(iv); or we can claim that internal standing beliefs meet some further condition(s) not on this list, which the notebook entries do not meet.

The first option is a difficult one. C&C devised the Otto case specifically to meet (i)–(iv); and even if there might be some doubts about how well the notebook entries satisfy each of them, such questions could easily be answered by modifying the example. For instance, C&C observe that it might be awkward to carry a notebook into the shower. But we can imagine that Otto carefully laminates the notebook's pages, or that he adopts some more high-tech solution. It seems clear, in any case, that an external device *could* contain information that meets (i)–(iv).

Now consider our second option for rejecting premise (2): to claim that internal standing beliefs meet some *other* condition, not included in (i)–(iv), that is necessary for being a standing belief, and that the notebook entries do not meet. I know of two candidates for such further conditions. The first, noted by C&C, is that there is a phenomenal difference between recalling an ordinary standing belief (e.g., that dinosaurs once roamed the earth), and looking up this information in a notebook. Ordinary recall has a kind of effortless immediacy: when the issue arises, I simply find myself occurrently thinking that dinosaurs roamed the earth. But it *feels* different to consult a notebook.

C&C acknowledge that there is some phenomenological difference between these, but they characterize this difference as 'shallow' for they deny that having a particular kind of phenomenological feel is necessary for being a standing belief. You might disagree, and see the phenomenological difference as deep and important. I will remain neutral on this issue for reasons that will become clear below. But suppose, for the moment, that there is a deep, important phenomenological difference between ordinary recall and consulting a notebook. In that case, we should add a fifth condition to our characterization of the functional role standing beliefs play.

(v) Recalling them has a particular phenomenology: roughly, the phenomenology of effortless immediacy.

Dan Weiskopf[7] suggests a second possible difference between internal standing beliefs and notebook records. Weiskopf points out that internal standing beliefs are automatically revised in light of new information. Here is a slightly modified version of an example he gives: when you learn that Sam and Max are married, you will probably come to believe that they live at the same address. If you then learn that they are no longer married, you will likely abandon the belief that they live at the same address. These further revisions in your beliefs are automatic, that is, they occur without deliberation. Weiskopf points out that Otto's notebook is not 'informationally integrated' in this way. When he writes in his notebook that Sam and Max are married, no entry reading 'Sam and Max share an address' automatically appears. Otto may, of course, write this entry as well—Weiskopf's point is just that this further revision requires an extra, deliberate step. Similarly, when Otto learns that MoMA has temporarily moved to Queens, the sentence 'MoMA is on 53rd Street' does not instantly vanish from the notebook.

Here, then, is another condition we might add to the original four:

(vi) They are informationally integrated.

Let us take stock. Ordinary internal standing beliefs meet conditions (v) and (vi) whereas the entries in Otto's notebook do not. Whether we should reject premise (2) for this reason depends on two questions. First, are (v) and (vi) necessary to being a standing belief? Second, granting that Otto's notebook entries do not satisfy (v) and (vi), might these conditions be satisfied by information stored in a different type of external device?

I suggest that the answer to the second question is 'yes' and that we can therefore ignore the first question. Suppose that, instead of a simple paper notebook, Otto carries an external computing device that is linked to his brain. In a lucid moment, or perhaps before he succumbs to full-blown Alzheimer's, Otto programs the device to constantly scan his thoughts and to perform as follows. When the device detects that Otto is thinking about dinosaurs, or about prehistoric times (etc.), it causes him to occurrently believe that dinosaurs roamed the earth. This process has the phenomenology of ordinary recall: the occurrent belief seems to come to him immediately and without effort. He also programs the device to be sensitive to additions or alterations of information. So when Otto learns that MoMA has moved to Queens, the device automatically adds or modifies any bits of stored information that are relevant to this new fact: e.g., it stores the information that MoMA is no longer on 53rd Street. Arguably, then, the information in this external device now meets conditions (v) and (vi).

Of course, the notebook example is much closer to the kinds of external information storage mechanisms that are in common usage today. So if (v) or (vi) is required for being a standing belief—an issue on which I'll remain neutral—then C&C have shown only that minds *can* be extended, not that they are *currently* extended. (Weiskopf recognizes this point, and argues only that minds are not, in fact, extended. He allows that they could *become* extended.) Still, the mere possibility of extended minds is all that's needed for what is perhaps C&C's central contention, that 'when it comes to belief, there is nothing sacred about skull and skin' (p. xx8xx).

There may be further conditions that are necessary for being a standing belief, which an external device could not meet. But none comes easily to mind. Nor do I expect one to emerge: for standing beliefs do seem to be defined, as such, by their causal relations to other states and processes, including occurrent ones. And there seems no principled reason to deny that something beyond the skin could play the same type of causal role as internal standing beliefs. So I suggest that it is reasonable to accept premise (2).

This leaves us with premise (4). Premise (4) may seem the least suspect of C&C's three major premises. But I will argue that, in fact, it is the most objectionable.

4. The Narrow Mind

Premise (4) is that Otto's standing beliefs are part of his mind. Now this may seem almost a definitional truth; after all, what is a belief except a mental state, and what is a mental state except part of a mind? But I will argue that, to avoid the consequences discussed in Section 2, we should deny that *standing* beliefs are part of the mind.

I think that C&C's examples are persuasive in illustrating the close affinity between internal standing beliefs and a variety of external states. In particular, I think that C&C make a compelling case for the following conditional.

> If standing beliefs are part of the mind, then the mind can be indefinitely extended: to notebooks, external computing devices, and even parts of others' minds.

I think that their parallel argument, regarding the affinity between processes performed by external devices and processes performed internally, also succeeds. It establishes this conditional:

> If nonconscious cognitive processes are part of the mind, then the mind can be indefinitely extended: to external computing devices and even parts of others' minds.

The consequents of both these conditionals—that the mind can extend to notebooks, external computing devices, and even parts of others' minds—have the worrisome consequences discussed in Section 2.

I suggest that we accept these conditionals, on the basis of C&C's ingenious arguments. But to avoid those worrisome consequences, we should reject their consequents and, hence, reject their antecedents. In other words, we should reject premise (4).

C&C have shown, I think, that internal standing states and nonconscious processes are essentially similar to states of notebooks and

external computational processes. From the fact that the latter, external states and processes are non-introspectible, we can infer that their internal equivalents are also non-introspectible. (There is independent reason to believe that standing states and nonconscious processes are nonintrospectible.[8]) And from the fact that behavior produced by external states and processes is not truly *action* (e.g., the robot's behavior, caused by Otto's external standing states and nonconscious processes, is not Otto's action), we can infer that behavior produced by internal standing states and nonconscious processes is not truly action.

The worrisome consequences that result from C&C's conclusion do not derive from the fact that the allegedly mental states and processes are *external*. Rather, they derive from the fact that the allegedly mental states and processes are *standing* states and *nonconscious* processes. Whether the states are internal or external is then unimportant; for, I think, C&C have shown that 'when it comes to belief, there is nothing sacred about skull and skin' (p. xx8xx).

The best option, then, is to reject premise (4). This means that the internal equivalents of notebook entries and external computing processes—namely, internal standing beliefs and nonconscious cognitive processes—are not, strictly speaking, part of the mind. On this view, the mind is made up entirely of *occurrent* states and *conscious* processes. These include beliefs or desires that are now being entertained, conscious thoughts, emotions, and sensations, and conscious cognitive processes.

C&C are well aware of the alternative I am suggesting. They recognize that 'it may be the most consistent way to deny Otto's belief' but reject it as 'extreme' (p. xx16xx).

> To consistently resist this conclusion [that Otto himself is an extended system], we would have to shrink the self into a mere bundle of occurrent states, severely threatening its deep psychological continuity [p. xx18xx].

This is a serious worry for my position. But it is worth noting that the problem is also faced by some familiar views about the mind, including Derek Parfit's bundle theory (chapter 70 of this volume).

I will close by briefly sketching how a mind, understood as a series of (sets of) occurrent states, can enjoy the psychological continuity we think ourselves to enjoy. The approach uses states that are not part of the mind as the causal ground for its psychological continuity.

Let us begin with an analogy. Consider an automobile factory—call it *Factory A*—that produces several different models of cars. The cars may bear the factory's trademark, '*A*' in which case they share an internal feature that reflects their common origin. Or they may be etched with consecutive numbers (e.g.,'7691' '7692'. . .) so that their internal features compose a recognizable pattern, regardless of whether there are more specific features they have in common. But even if they bear no such internal marks, they still form a unified, causally salient class, viz., the class 'Products of Factory A' The factory's causal continuity grounds the causal salience of that class, as products of the same cause.

For instance, suppose that the machine that produces disc brakes is gradually deteriorating. The deterioration leads each disc brake it produces to be a bit less round—slightly further from perfect roundness—than the last. Now while the machine itself does not belong to the class 'Products of Factory A' the cars that are part of that class are causally unified in that they are all products of a common cause. The decreasing roundness of their brakes, due to the ongoing deterioration of the machine, is one illustration of this causal unification. The deterioration merely reflects the causal unity: the cars would be causally unified, in the sense that I intend, even if the machines were always in ideal working order.

A similar picture applies to occurrent states. (For simplicity here, I'll assume materialism about standing states; the picture could be adjusted to apply to dualism about such states.) Suppose that a number of occurrent states have a shared origin. They spring from states, including standing beliefs and non-mental states, of a persisting physical organism, perhaps together with salient features of its environment. This shared origin may produce occurrent states that share internal features: e.g., one's occurrent states may share a quality of hopefulness because they spring from physically-based dispositions to be optimistic. Or they may have complex rational interrelations that depend on non-mental states: e.g., yesterday's occurrent belief 'It will rain tomorrow' may cause a standing belief to that effect, which the next day causes the occurrent belief 'it will rain today'

But as in the factory case, it is the causal continuity of the shared origin itself that renders the states causally unified. Any shared internal features or coherent rational structure is unnecessary—for instance, it is not present

in the presumably jumbled occurrent states of Alzheimer's patients like Otto. And the shared origin needn't be included in the class of occurrent states in order to ground the class's causal unity, any more than the factory needs to be included in the set of cars to ground that set's causal unity.

How much psychological continuity do we possess? Some of us are psychologically stable, consistent, and predictable. These steady types experience occurrent states that follow previous occurrent states in fairly predictable ways. (*Modulo* the continuity of external stimuli, of course.) So the claim that the mind or self is constituted by occurrent states does not raise the specter of radical discontinuity, when applied to these stable individuals.

The more problematic cases are those individuals who undergo a series of radically disparate occurrent states. Taking these unstable minds to be constituted by the series of sets of occurrent states seems to threaten psychological continuity, just as C&C allege. But I don't think that this is a problem. First, I think that it makes sense to deny that unstable characters enjoy one sort of psychological continuity, viz., the continuity involved in one's experiences and thoughts *seeming*, to one, to have a rational structure. The lack of apparent rational structure among occurrent experiences and thoughts is, after all, why dementia and Alzheimer's disease leave their victims so bewildered.

But there is another sense of 'psychological continuity' that even the least stable individuals possess. This is the continuity that is responsible for the fact that the succession of wildly varying, seemingly unrelated states is fully *explainable*. The factors that explain it are underlying physical states and processes. The point of contention between my view and C&C's concerns whether these underlying factors—which include standing states and nonconscious processes—are themselves *part of* the mind. I have argued that, in a strict, principled sense of 'mental' they are not. They are outside the realm of introspectibility; and including them within the mind will extend the range of a subject's actions to staggering proportions. There are therefore strong reasons to deny that merely standing states and nonconscious process are, strictly speaking, part of the mind.

Conclusion

Obviously, many questions remain. But I hope to have shown that a serious alternative to extending the mind is to reject premise (4) of C&C's argument, and to limit the mind to occurrent, conscious states and processes. In fact, I think that this fits with their claim that 'when it comes to belief, there is nothing sacred about skull and skin.' I fully agree with this conclusion, but I draw a different moral from it. They conclude that some external states and processes are mental; I conclude that some internal (standing) beliefs and (nonconscious) cognitive processes are non-mental.

It is surprising to think that standing beliefs and nonconscious processes lie outside the mind, even if they are inside the brain. Still, on balance, this conclusion seems less costly to intuitions and hence ultimately more credible than the claim that our mind can extend to notebooks, external computing devices, and others' minds.[9]

NOTES

1. All page numbers will refer to the reprint in this volume.
2. Strictly speaking, the argument requires only that the notebook entries at least *partially constitute* Otto's standing beliefs. For whatever partially constitutes a part of the mind is itself a part of the mind.
3. Of course, the upshot of the imagining—the visualization of the rotated shape—is conscious. But the *process* of rotating it is, at least in their example, nonconscious.
4. Elsewhere, I provide a more detailed argument to show that these wide states do not crucially contribute to action. (Gertler, "The Narrow Mind," in preparation.)
5. We can also suppose that the robot could have awakened the organism, in case of an emergency—so the organism was only partially and temporarily unreceptive to input from the robot. But this may not matter, since Otto does not consult his notebook when he is in a deep sleep either.
6. See p. xx17xx. C&C give these conditions as a rough outline of the relevant functional role. They do not claim that they are jointly sufficient for being a standing belief, or that each is necessary. In fact, they express some doubts as to whether (iv) is necessary. These details will not affect my argument.
7. Weiskopf, "Patrolling the Mind's Boundaries," (MS, 2006)
8. To use a historical example: Descartes' meditator can introspect his occurrent beliefs that 'I doubt that I am sitting before the fire' or '2+3=5' but cannot introspect the causal sources of those beliefs, including standing beliefs or past cognitive processes. This is why he cannot rule out, through introspection

alone, the possibility that these beliefs are caused by an evil genius, rather than by a standing belief or a past (and hence currently nonconscious) cognitive process.

9. I am indebted to the participants in the National Endowment for the Humanities Institute on Consciousness and Intentionality in Santa Cruz, where I presented an ancestor of this paper in July 2002—especially Terry Horgan, Amy Kind, Eric Schwitzgebel, Galen Strawson, and Aaron Zimmerman. I am also deeply grateful to Dave Chalmers, for helpful discussion and for correcting several errors. Larry Shapiro provided extensive, valuable comments on this version of the paper.

The Embodied Mind

Shaun Gallagher and Dan Zahavi

Let's start with a fully cognizing human being who is complete in body and mind, and ask what we could subtract while still retaining a cognizing mind. Such thought experiments may help us to home in on precisely what a cognitive system or mind actually is.

Let's take as our example any one of us who happens to be complete in body and mind, understanding that in an ordinary and everyday way. Of course, some of us may not be bodily complete. One of you who is holding this book in your hand right now may be missing the other hand, perhaps as a result of an accident and amputation. In any case, let's take as our example someone with all her limbs. Now we can ask, would it make any difference in regard to her ability to think or imagine or remember, or to engage in most cognitive exercises if she were missing one or even several of her limbs? Right, it doesn't seem that we need all our limbs to engage in cognizing. So let's get rid of them. While we're at it, we may as well get rid of all the other extraneous body parts—those that we don't seem to need in order to do our thinking. As these kinds of thought experiments go, we usually end up with just our brains, since even sensory input can be provided artificially. For example, we can directly stimulate the parts of the brain responsible for registering sensory information and thereby, supposedly, have exactly the same experience we would have if our sensory organs were delivering that information. This common thought experiment is referred to as the brain-in-the-vat, and the image is of a brain floating around in a vat of chemicals, kept alive by artificial nourishment, and kept informed by various electrodes that carry information about the world, or about whatever the mad scientist running this experiment wants to feed it.

Dennett 1981 has taken this thought experiment one step further. He tells the story of being sent on a mission that involved removing his brain, storing it in a vat, but remaining connected with the body via radio waves. Dennett's mission, however, is a dangerous one, and in the process his body ceases its biological functioning and, in effect, dies. His brain, however, is still alive in the vat. Understandably, he gets upset:

> Waves of panic and even nausea swept over me, made all the more horrible by the absence of their normal body-dependent phenomenology. No adrenaline rush of tingles in the arms, no pounding heart, no premonitory salivation. I did feel a dread sinking feeling in my bowels at one point, and this tricked me momentarily into the false hope that I was undergoing a reversal of the process that landed me in this fix—a gradual undisembodiment. But the isolation and uniqueness of that twinge soon convinced me that it was simply the first of a plague of phantom body hallucinations that I, like any other amputee, would be all too likely to suffer. (1981, p. 225)

As time goes on, Dennett is provided with a new body, which he finds difficult to master, but which, after a period of adjustment, seems just fine. He thinks perhaps this is similar to undergoing extensive plastic surgery or a sex-change operation. Dennett subsequently learns, however, that the technicians had copied his

Excerpted from S. Gallagher and D. Zahavi, *The Phenomenological Mind* (second edition) (Routledge, 2012). Reprinted by permission of the publisher.

brain's functional structure and all of the information in it to a computer program, and that he is doubly connected—to his brain, and, with the flick of a toggle switch, to the computer running his artificial brain. He is able to flip a switch between brain and computer, but is unable to tell the difference in his experience.

The moral of this story, if we were to follow Dennett's line of reasoning and put this story together with the standard line of thought about brains-in-vats, is that not only is the body unnecessary for experience and cognition, but we don't even need the brain, as long as we have the program and information running on the right kind of hardware. This constitutes a functionalist perspective according to which artificial neural-net processing of information can generate the same mental experiences as can be generated by the brain alone. What's important is not the physical instantiation (although this is certainly a consideration, since it would likely make a difference if we tried to run the software on a Mac or a PC rather than on a sophisticated neural net computer); rather, what is important is the software program and the information that constitutes the essential part of a system required to generate me, and my cognitive life. Once we have the right information and the proper brain-replicating syntax, we should be able to generate your cognitive experience in any machine that can run the program.

Does this mean that the body contributes nothing of importance to the cognitive system? Of course, we can say that the brain is important in all normal cases where we do not have a backup artificial brain. And even Dennett suggests that to *do* anything, to take action in some way, one might need some kind of body. But we could think that a robotic body could do just as well, as long as it were properly connected (by radio transmitters) to the artificial brain.

The image of the brain-in-the-vat is surprisingly influential even for opponents of functionalism. Thus, Searle, who takes an anti-functionalist view, and emphasizes the importance of neurobiology, nonetheless, in defending a radical form of internalism, appeals to the same image:

> Even if I am a brain in a vat—that is, even if all of my perceptions and actions in the world are hallucinations, and the conditions of satisfaction of all my externally referring Intentional states are, in fact, unsatisfied—nonetheless, I do have the Intentional content that I have, and thus I necessarily have exactly the same Background that I would have if I were not a brain in a vat and had that particular Intentional content. *That* I have a certain set of Intentional states and *that* I have a Background do not logically require that I be in fact in certain relations to the world around me. . . .
>
> (Searle 1983, p. 154)

This kind of denial of the cognitive significance of the body has a long tradition. Compare the following statement in Plato's dialogue, *Phaedo*:

> It seems that so long as we are alive, we shall continue closest to knowledge if we avoid as much as we can all contact and association with the body, except when they are absolute necessary, and instead of allowing ourselves to become infected with its nature, purify ourselves from it until God himself gives us deliverance.
>
> (Plato 1985, 67a)

Such a disembodied view on the mind was also found in classical cognitive science, since it examined intelligent behaviour as if it were independent of any specific bodily form. Indeed, until recently, insofar as neuroscientists considered the body, it was only qua its representation in the somatosensory cortex.

Now one might think that it is incumbent on the phenomenologist, or on the theorists of embodied cognition, to show that there is no cognition without embodiment. But there are two questions here. First, the in-principle question of whether the notion of a disembodied brain (a brain in a vat) is at all intelligible; and second, regardless of what the answer is to the first question, we can ask whether human cognition de facto is disembodied. We can answer the second question (in the negative) without having to demonstrate that the brain-in-the-vat thought experiment is an unintelligible thought experiment. It just is an empirical fact that we are indeed embodied, that our perceptions and actions depend on the fact that we have bodies, and that cognition is shaped by our bodily existence. This is, we might say, a 'no-brainer.' But we can note the following in response to the first question. The brain-in-the-vat thought experiment actually shows that perception and action do require some kind of embodiment. Even the pure brain-in-the-vat requires absolutely everything that the body normally provides—for example, sensory input and life support. Indeed, the importance of the body can be measured in considering precisely what it would take to sustain a disembodied brain and the supposed experience that goes along with it. What is possible for a brain-in-the-vat is only possible if it

is provided with a properly balanced nutrition, a properly balanced mix of hormones and neurotransmitters, and a complex stream of sensory information, properly adjusted for the temporal differentiations that are in fact involved in intermodal binding. If we consider only the visual input, we would have to assume that any poking around in the visual cortex that would replicate our human visual experience would have to be so specified in its details, that an analogue or digital input mechanism would have to be as complicated, as chemically complex, and as enactive as the human eye. That is, the full and extraordinary support system that would be required to allow a brain-in-a-vat to experience things as we experience them, or in other words, to allow a brain-in-a-vat to be phenomenologically in-the-world and not just physically in-a-vat, would have to replicate the bodily system that already supports our ordinary existence.

Whether or not a brain-in-the-vat is a real possibility, it is certain that *our* cognitive experience is shaped by an embodied brain. Indeed it is increasingly accepted that the brains we have are shaped by the bodies we have, and by our real-world actions. Cognition is not only embodied, it is situated and, of course, it is situated because it is embodied.

The fact that we stand upright is distinctive for the human species, and this biological fact, which comes along with many other biological facts, has far-reaching consequences with respect to perceptual and action abilities, and by implication, with respect to our entire cognitive life. Erwin Straus, for example, points out that 'the shape and function of the human body are determined in almost every detail by, and for, the upright posture' (1966, p. 138). Consider a brief list:

- First, in regard to *human anatomy and skeletal structure*, the upright posture requires a specific shape and structure of the human foot, ankle, knee, hip, and vertebral column, as well as the proportions of limbs, and all of this demands a specific musculature and nervous system design. In terms of evolution, the shaping of the body for the upright posture also permits the specifically human development of shoulders, arms, hands, skull, and face. The important point here is that these anatomical structures define our capabilities and therefore define what counts as the world. Gibson 1986 developed the idea that objects in the environment can afford different kinds of action, given the kind of body that we have. Such affordances are closely tied to our bodily shape and our action capabilities. A chair affords sitting precisely because the human body bends at the knees, etc. Capabilities to sit or to adopt some other posture are first of all motor; but they extend to the most abstract and rational capacities for cognition, such as counting and the development of mathematics (see Johnson 1987, 2007; Lakoff and Johnson 1980; Lakoff and Nuñez 2001; Sheets-Johnstone 1990).
- Second, in terms of *development*, attaining the upright posture is delayed in humans. The infant is required to learn it in a struggle against gravity. This calls for a basic conscious wakefulness: if you fall asleep, you fall. Posture and movement are directly related to biological states of sleep and wakefulness. Prior to standing, early crawling behaviour influences the development of perception and cognition (Campos et al. 1992). The change of posture that comes with standing and walking equally affects what we can see, what we can attend, what we can grasp and manipulate.
- Third, in terms of *how we are related to things and other people*, with the upright posture we maintain distance and independence—distance from the ground; distance from things; and some degree of independence from other people. In standing, the range of vision is extended, and accordingly, the environmental horizon is widened and distanced. The spatial frameworks for perception and action are redefined. Standing frees the hands for reaching, grasping, manipulating, carrying, using tools, and pointing. Both phylogenetically (with respect to evolution) and ontogenetically (with respect to individual development), these changes introduce complexities into brain structure, complexities that eventually help generate rational thought (Paillard 2000).
- With respect to our *perceptual abilities*, which in turn shape all other cognitive abilities, in evolutionary terms, attaining the upright posture means that the olfactory sense declines in importance; seeing becomes primary. We are able to see far ahead of where we are currently located, and this grants foresight and allows for planning. While our hands are liberated for

> more proficient grasping and catching, our mouths are liberated for other purposes, since olfactory mechanisms (required for finding our way around when close to the ground) shrink and no longer dominate facial structure. The jaw structure not only defines what we eat, but along with the development of the more subtle phonetic muscles, enables the development of vocal language. And if you ask Aristotle, he'll tell you that this means the development of both politics and rationality.

Add to this the idea that the body 'preprocesses' and filters incoming sensory signals, and 'postprocesses' and limits efferent signals that contribute to motor control. Comparative anatomy shows that the shape and relative locations of the ears, for example, allow us to determine the direction of a sound (Chiel and Beer 1997). Bodily movements are not fully determined at brain level; rather, they are reengineered by the design and flexibility of muscles and tendons, their geometric relations to other muscles and joints, and the prior history of their activation (Zajac 1993). Thus, 'the nervous system cannot process information that is not transduced by the periphery, nor can it command movements that are physically impossible for that periphery' (Chiel and Beer 1997, p. 554). These observations are part of a larger story about what shapes the body and how that also shapes cognition. But they are sufficient to indicate that the biological body (what it enables and excludes by its structure, basic posture, and motor capacity) is the body that shapes the way that we perceive and think about the world.

Robotic and Biological Brains

These are lessons hard won, not by means of thought experiments, where you seemingly can think many aspects of the system away, but in the 'real world' of the biological sciences, as well as in advanced robotics. Roboticists, like Rodney Brooks at MIT, have discovered that the traditional approaches of trying to develop robots from the top down, i.e. starting with a disembodied syntax and trying to add a functional artificial body that would heed the commands from a central intelligence computer, just don't work. Rather, their more recent initiatives are attempts to design robots from the bottom up, building simple, pragmatically-ordered, biologically-inspired, sensorimotor machines that can move around environments by using information gathered in real time from the environments themselves. Such robots are 'physically grounded,' which means they are physical entities embedded in their physical environments, but in a way in which their representations pragmatically reference the real world; they are enactive perceiving machines, grasping the world in terms of projects to be accomplished.

> Nouvelle AI is based on the physical grounding hypothesis. This hypothesis states that to build a system that is intelligent it is necessary to have its representations grounded in the physical world. Our experience with this approach is that once this commitment is made, the need for traditional symbolic representations soon fades entirely. The key observation is that the world is its own best model. It is always exactly up to date. It always contains every detail there is to be known. The trick is to sense it appropriately and often enough. To build a system based on the physical grounding hypothesis it is necessary to connect it to the world via a set of sensors and actuators.
>
> (Brooks 1990, p. 5)

Brooks comes to this important realization, however, not by thinking of robots as bodies, but by thinking of bodies as robots:

> The body, this mass of biomolecules, is a machine that acts according to a set of specifiable rules. . . . We are machines, as are our spouses, our children, and our dogs. . . . I believe myself and my children all to be mere machines.
>
> (Brooks 2002, pp. 173–75)

The philosophical background to this particular way of conceiving of the body is clearly Cartesian. Descartes characterized animals as purely physical automata—robots devoid of consciousness. This was extended to humans by a variety of philosophers, including La Mettrie 1745 and Cabanis 1802, and was further explicated by Hodgson 1870 and Huxley 1874. Brooks seems bound to follow in this tradition, perhaps topping it off by proposing that conscious-like intelligence should emerge from this kind of system. An alternative philosophical backdrop to the concept of embodied cognition, however, is alive and well. This is worked out in the phenomenological views of Husserl and Merleau-Ponty, and updated by philosophers and scientists like Clark 1997; Varela et al. 1991; Thompson 2007; Thompson and Varela 2001; Sheets-Johnstone 1990, 1999; Michael Wheeler 2005; and others. This

alternative approach follows Merleau-Ponty in rejecting the idea that the body is simply a 'highly polished machine' (1962, p. 76). So let's dig deeper into the meaning of embodiment, how it situates us and how it shapes our cognitive experience.

Phenomenology of the body—a very short history

The best-known philosopher of embodiment is undoubtedly Merleau-Ponty. But Merleau-Ponty was certainly not the only phenomenologist who devoted time and energy to a painstaking analysis of the lived body. Not only are there other French phenomenologists who have written extensively on the body, Sartre and Michel Henry, for example, but it would also be a mistake to identify phenomenology of embodiment with French phenomenology. Already in Husserl's lecture course *Thing and Space* from 1907 one can find far-reaching phenomenological analyses of the moving and sensing body. And it is well known that Husserl's analysis of the body in the second volume of his *Ideas* served as a decisive inspiration for Merleau-Ponty's *Phenomenology of Perception.* Husserl's manuscript was only published posthumously in 1952, but Merleau-Ponty visited the Husserl archives shortly before the outbreak of the Second World War—as one of its first foreign visitors—and had on that occasion a chance to read Husserl's unpublished manuscript (cf. Zahavi 1994, 2006). But even Husserl might not have been the first. Michel Henry has argued that one can find an implicit theory about the lived body in the most famous dualist of them all, namely Descartes (Henry 1975, p. 139). And if one moves forward in history, to the Napoleonic wars, one will encounter the work of another French philosopher Maine de Biran who—and this is still according to Michel Henry—provides a phenomenological account of the body that is superior to the ones subsequently to be found in the writings of Husserl, Sartre, and Merleau-Ponty.

The phenomenological investigation of the body is not the analysis of one object among others. That is, it is not as if phenomenology in its investigation of a number of different ontological regions (the domain of logic, mathematical entities, utensils, works of art, etc.) also stumbles upon the body and then subjects it to a close scrutiny. On the contrary, the body is considered a constitutive or transcendental principle, precisely because it is involved in the very possibility of experience. It is deeply implicated in our relation to the world, in our relation to others, and in our self-relation, and its analysis consequently proves crucial for our understanding of the mind–world relation, for our understanding of the relation between self and other, and for our understanding of the mind-body relation.

The phenomenological emphasis on the body obviously entails a rejection of Cartesian mind-body dualism. But it should be just as obvious that this does not entail an endorsement of some kind of Cartesian materialism. It is not as if the phenomenological way to 'overcome' dualism is by retaining the distinction between mind and body, and then simply getting rid of the mind. Rather, the notion of embodiment, the notion of an embodied mind or a minded body, is meant to replace the ordinary notions of mind and body, both of which are derivations and abstractions. Merleau-Ponty famously speaks of the ambiguous nature of the body, and argues that bodily existence is a third category beyond the merely physiological and the merely psychological (1962, p. 350). The lived body is neither spirit nor nature, neither soul nor body, neither inner nor outer, neither subject nor object. All of these contraposed categories are derivations of something more basic.

Phenomenologists object to the metaphysical division between *res extensa* and *res cogitans.* If one accepted such a division, the only place for the body would seem to be on the side of the *res extensa.* But phenomenologists deny that the body is a mere object in the world. The body is not merely an object of experience that we see, touch, smell, etc. Rather, the body is also a principle of experience, it is that which permits us to see, touch, and smell, etc. Obviously, the body can also explore itself. It can take itself (or the body of another) as its object of exploration. This is what typically happens in physiology or neurology, etc. But such an investigation of the body as an object is not exhaustive. As Sartre famously points out, we should be careful not to let our understanding of the lived body be determined by an external perspective that ultimately has its origin in the anatomical study of the *corpse* (1956, p. 348; cf. Merleau-Ponty 1962, p. 351). As he continues in *Being and Nothingness*:

> The problem of the body and its relations with consciousness is often obscured by the fact that while the body is from the start posited as a certain *thing* having its own laws and capable of being defined from outside, consciousness

> is then reached by the type of inner intuition which is peculiar to it. Actually if after grasping 'my' consciousness in its absolute interiority and by a series of reflective acts, I then seek to unite it with a certain living object composed of a nervous system, a brain, glands, digestive, respiratory, and circulatory organs whose very matter is capable of being analyzed chemically into atoms of hydrogen, carbon, nitrogen, phosphorus, etc., then I am going to encounter insurmountable difficulties. But these difficulties all stem from the fact that I try to unite my consciousness not with *my* body but with the body of *others*. In fact the body which I have just described is not *my* body such as it is *for me*.
>
> (Sartre 1956, p. 303)

The phenomenological contribution to a solution of the mind-body problem does not take the form of a metaphysical theory of mental causation, nor does it consist in an explanation of how the body interacts with the mind; rather, it seeks to understand to what extent our experience of the world, our experience of self and our experience of others are formed by and influenced by our embodiment. But through this change of focus, it also rethinks and questions some of the distinctions that define the mind-body problem in the first place.

The first and most basic phenomenological distinction to be made, and the one that allows us to see that Brooks may be working with the wrong concept of the body, is between the *objective body* and the *lived body* (Husserl's distinction between *Köper* and *Leib*, respectively; Merleau-Ponty's distinction between the *Le corps objectif* and the *corps proper* or *corps vécu*). This is a phenomenological distinction rather than an ontological one. It is not meant to imply that each of us has two bodies: one objective and one lived. Rather, it is meant to explicate two different ways that we can experience and understand the body (Husserl 1973a, p. 57). Whereas the latter notion captures the body understood as an embodied first-person perspective, the former focuses on the body as seen from an observer's point of view, where the observer may be a scientist, a physician, or even the embodied subject herself. I can view my own body as if from the outside. I can look at my hand and think, 'Hmm, how truly odd that this thing has five wiggling digits.' The objective body is, in varying degrees of abstraction, and defined in a variety of perspectives (neurological, physiological, anatomical), a perceived body; it is the objectification of a body which is also, nonetheless, lived. Looking at the body as a thing that can be analysed, dissected, objectively understood, in the way that we might understand a machine or a robot, is clearly important for making progress in the biological sciences, in medicine, and perhaps in robotics. If we are taking this perspective on the body, we are taking a third-person perspective—examining the body as something that we, as subjects, can observe as an object.

In contrast, of course, the only way we can make such observations, or any observations, is if we are in fact an experiencing, sensorimotor, living body—if we have eyes that see, hands that are capable of haptic touch, ears that hear, and so forth. In this regard I do not observe or contemplate my hand, I reach out with it and grab something. The body as subject, as experiencer, as agent, rather than the body as object, as thing experienced—this is a basic distinction missed by the Cartesian tradition. When Descartes—according to the standard interpretation—insists that he is a thinking thing, and is not his body, which is an extended thing, he thinks that he can think without his body thinking. In fact, however, Descartes was able to think such thoughts only because he was a living body that included a highly inter- and intra-connected brain. As far as we can tell, Descartes stopped thinking these thoughts in the early morning of 11 February 1650 when he died. An autopsy on his objective body would have shown a serious respiratory infection as cause of death.

A description of the lived body is a description of the body from the phenomenological perspective. On the one hand, it is the way the body appears in experience. On the other hand, it is much more than that—it is the way the body structures our experience. The body is not a screen between me and the world; rather, it shapes our primary way of being-in-the-world. This is also why we cannot first explore the body by itself and then subsequently examine it in its relation to the world. On the contrary, the body is already in-the-world, and the world is given to us as bodily revealed (Husserl 1971/1980, p. 128). Indeed, as Sartre points out, the body is operative in every perception and in every action. It constitutes our point of view and our point of departure (1956, p. 326):

> The case could not be otherwise, for my being has no other way of entering into contact with the world except to *be in the world*. It would be impossible for me to realize a world in which I was not and which would be for me a pure object of a surveying contemplation. But on the contrary it is necessary that I lose myself in the

> world in order for the world to exist and for me to be able to transcend it. Thus to say that I have entered into the world, 'come to the world,' or that there is a world, or that I have a body is one and the same thing.
>
> (Sartre 1956, p. 318; cf. Merleau-Ponty 1962, p. 82)[1]

We have a sense of the body in what it accomplishes. I have a tacit sense of the space that I am in (whether it is crowded, whether it is wide open, or whether it is closing in). Likewise, I have a proprioceptive sense of whether I am sitting or standing, stretching or contracting my muscles. Of course, these postural and positional senses of where and how the body is tend to remain in the background of my awareness; they are tacit, recessive. They are what phenomenologists call a 'pre-reflective sense of myself as embodied.'[2]

This sense of embodiment is not simply spatial. I can feel sluggish (after eating a heavy dinner, for example) or clumsy; or I can feel energetic and fully attuned to my surroundings (after exercising or yoga, for example). If I am depressed by some bad news, I can feel that in my body; if I am elated by good news or buoyed up by an impending challenge—these are feelings and moods that I feel bodily. If I am angry or fearful or happy and comfortable, these are emotions that I feel bodily. Moreover, all of these aspects of embodiment shape the way I perceive the world. If I'm depressed, the world seems depressing; if I am elated, the world seems promising; if I am hungry, as William James noted, an apple appears larger than when I am satiated. Since this is the lived body with which I perceive and act, it is in constant connection with the world. And this connection is not a mere surface-to-surface contact, as a corpse might lie on the surface of a table; rather, my body is integrated with the world. To be situated in the world means not simply to be located someplace in a physical environment, but to be in rapport with circumstances that are bodily meaningful. It means something if the drink that I want is out of reach; if I am unable to sprint as fast as I need to when I am being chased by a ferocious animal, or in danger of being run down by a bus. Those possibilities that my body enables, and that define the environment as a world of affordances, just as much as those activities that my body prevents or limits, and that define what is possible or impossible—these are aspects of embodiment that I live with, and through, and that define the environment as situations of meaning and circumstances for action.

Much more could be said about the body-environment relation, since the environment is not simply a place where we perform our actions. The environment directly and indirectly regulates the body, so that the body is in some sense the expression or reflection of the environment. The environment calls forth a specific body-style so that the body works with the environment and is included in it. The posture that the body adopts in a situation is its way of responding to the environment. The body finds itself already with feelings, drive-states, kinaesthetic sensations, etc. and they are partially defined by the environment in which it must function. The 'internal environment' of the body, which functions homeostatically and automatically, and is constituted by innumerable physiological and neurological events, is simply an internalized translation and continuation of the 'external' environment. Changes in the 'external' environment are always accompanied by changes in the 'internal' one, e.g. 'changes induced in the blood by alterations in the [external] environment, such as increased carbon dioxide or decreased oxygen tension in the inhaled air, and alterations in the temperature of the environment are minimized by appropriate alterations in circulation, respiration, and endocrine activity' (Gellhorn 1943, p. 15). All of these automatic regulations take place and are lived in bodily performances that are subpersonal and anonymous, although the results of this anonymous living are surely reflected, directly or indirectly, in the experience of the subject. It is also the case that when there are changes in the 'internal' environment, the 'external' environment can suddenly take on a different significance—i.e. the environment can become experientially different. The onset of eyestrain is a good example, as is the phenomenon of hallucination (see Gallagher 1986).

Nothing in this conception of embodiment should lead us to conceive of the body as something static, as if it has a fixed set of skills and abilities. The situation is quite different. Not only can the body expand its sensorimotor repertoire by acquiring new skills and habits, it can even extend its capacities by incorporating artificial organs and parts of its environment (Leder 1990, p. 30). In acquiring new skills, for example, we may begin by paying close attention to certain rules of performance, and when doing so we typically focus on and monitor our own bodily performance to an unusually high

degree. But a successful acquisition of this new ability will lead to performance without explicit monitoring of bodily movement; the skill becomes fully embodied and embedded within the proper context. As Leder has pointed out, 'A skill is finally and fully learned when something that once was extrinsic, grasped only through explicit rules or examples, now comes to pervade my own corporeality. My arms know how to swim, my mouth can at last speak the language. . . . A skill has been incorporated into my bodily 'I can' (1990, p. 31). This process of incorporation also has a marked temporal significance. Practice makes perfect because it habitualizes the skill. That which is practised in the past becomes embedded in my present bodily repertoire, and allows me to cope differently with new arising situations.

It is also possible to extend the capacities of the lived body by means of artificial extensions. Or to put it differently and perhaps even more strikingly, the lived body extends beyond the limits of the biological body. It doesn't stop at the skin. The classical example is the blind man's cane (a frequent example in the literature since Head (1920) first mentioned it). When first employing such a cane one experiences it as an external object exerting an impact upon the hand. But as the tool is mastered, one begins to feel through it to the experiential field it discloses (Leder 1990, p. 33). As Merleau-Ponty writes: 'The blind man's stick has ceased to be an object for him, and is no longer perceived for itself; its point has become an area of sensitivity, extending the scope and active radius of touch, and providing a parallel to sight' (1962, p. 143). Something similar can happen with the use of far more complex technologies. Consider, for example, the well-known experiments of Bach-y-Rita with sensory substitution and the technology known as Tactile Vision Sensory Substitution (TVSS) (see Bach-y-Rita et al. 1969, 2003; Gonzalez and Bach-y-Rita 2003). The TVSS was designed to provide vision to blind subjects. It maps images from a video camera to a vibrotactile belt worn on the back or abdomen. Because of the intermodal nature of sensory perception, we can, with some learning, 'see' the environment using tactile or auditory prostheses. The stimulation of the skin generates a quasi-visual experience of the environment. In recent development of this technology similar experience is generated by an electro-tactile tongue display unit. The intermodal sensory system of the body translates tactile signals on the skin into something like a visual experience of the external environment. Once the subject is habituated to the tactile stimulation the technology itself ceases to be an object and is incorporated into the body in a way that discloses the world. Such technologies, which are clearly objective pieces of engineering, can capitalize on sensorimotor contingencies and brain plasticity and become part of the body that we live.

Cole et al. 2000 provide another example by describing a virtual reality setup that links a human agent to a NASA robot, allowing the agent to steer the robot's arms by moving his own, and to see the robot's visual field through cameras mounted in the robot's head. After a few minutes of practice with this technology, the agent starts to have a strong sense of embodiment with the robot. As far as we know, and this is in contrast to the claim that our bodies are simply robots (Brooks), this would not work the other way around.

Thus, I can extend my set of skills and potential actions. I can do this through practice (as we can see in dance, athletics, etc.) or through artificial augmentation (as we see in sensory substitution technologies or robotics). What we describe as the lived body from the phenomenological perspective is exactly the same body as the biological body that we study from an objective perspective. The lived body clearly has a physiological basis, and as such it can be defined as 'a certain power of action within the framework of the anatomical apparatus' (Merleau-Ponty 1962, p. 109). Accordingly, it can suffer losses as well as experience gains. Thus brain lesions can occasion various forms of bodily self-alienation. One example is provided by the condition known as *anosognosia for hemiplegia.* Many right-hemisphere stroke patients deny their left-side paralysis. This denial typically remains despite the manifest demonstration of the paralysis. In one case, a patient with left paralysis claimed that she could walk, could touch the doctor's nose with her left hand, and could clap, when instead all she was doing was making motions only with her right hand (Ramachandran and Blakeslee 1998). When pressed, the patients may venture completely out of the realm of reality in defending their ability to move, stating that the immobile limb belongs to someone else, or is not a limb at all. One famous story tells of a patient who claimed that his paralyzed hand belonged to the doctor. When the doctor showed the patient his own two hands and asked how it was possible that he should have three hands, the

patient calmly replied, 'A hand is the extremity of an arm. Since you have three arms, it follows that you must have three hands' (Bisiach 1988, p. 469).

A stroke in the right hemisphere of the brain might result in symptoms of unilateral neglect. Patients fail to attend to the left side of their own body, or to respond to stimuli, objects, and even people located to their left side. This neglect finds astonishing manifestations. For instance, when served food, the patients will only eat the food that is on the right side of the plate, and will then complain that the hospital is starving them, and that they are not given enough to eat. If asked to copy a drawing, they will only copy one half of it. Furthermore, it was recently shown that this neglect not only affects our visual power, but also our power of imagination and recollection, thereby stressing the intricate interplay between these different forms of intentionality. During a two-minute period, patients were asked to mention as many French cities as they could remember. If these city names were subsequently marked on a map, it was discovered that all the mentioned cities were located in the east of France. No cities in the west (or left) of France were mentioned. In another experiment, patients from Milan were asked to think of the Piazza del Duomo, a place they knew very well. They were asked to imagine first that they were standing at the steps of the cathedral looking away from it, and they were then asked to describe what they were visualizing. They would only describe the right side of the square. They were then asked imaginatively to move to the opposite side of the piazza. When asked to describe what they were visualizing when they faced the cathedral, they would still only be describing the right side of the square. But, of course, what this means is that they were now describing the parts of the square that they had 'forgotten' a moment ago, whereas the parts of the square that they had just described were now lost to them (Bisiach and Luzzatti 1978).

These few references to pathology point to the important insight that core features of subjectivity can be sharply illuminated through a study of their pathological distortions. Pathological cases can function heuristically to make manifest what is normally simply taken for granted. They serve as a means of gaining distance from the familiar, in order better to explicate it. This is something phenomenology has long insisted upon, and it is no coincidence that especially the area of clinical psychopathology has attracted a lot of attention from phenomeno-logists, and that there exists a long-standing tradition of phenomenological psychiatry in France and Germany. Important figures include Minkowski, Binswanger, Tatossian, Tellenbach, and Blankenburg (cf. Parnas and Zahavi 2002; Parnas et al. 2002).

Thus, we can best come to understand our form of embodied life as it exists for us not in hypothetical or far-fetched thought experiments, but precisely in the ordinary cases of habit formation, and in the extraordinary cases of sensory substitution and pathological loss. To understand how the lived body works and how it shapes our cognition, we may be able to use high technology and robotics, but we clearly need both phenomenology and biology.

How the Body Defines the Space of Experience

One influential conception of knowledge takes knowledge to be a question of faithfully mirroring a mind-independent reality. If we want to know true reality, we should aim at describing the way the world is, not just independently of its being believed to be that way, but independently of all the ways in which it happens to present itself to us human beings. What we have been suggesting, however, is that this goal is illusory and unattainable. Even when doing science we have to start from an embodied perspective that we never fully escape. As Merleau-Ponty 1962 (p. 67) puts it, in response to a similar proposal for attaining a view from nowhere made by Leibniz, 'Is not to see always to see from somewhere?' This is a thought that is fully consistent with embodied and situated perception.

The 'somewhere' is a zero-point set by the perceiving body. Out of it a perspectival spatiality opens up. Although the objective body can be given a position in this perspectival space, the lived body cannot. 'The outline of my body is a frontier which ordinary spatial relations do not cross. This is because its parts are inter-related in a peculiar way: they are not spread out side by side, but enveloped in each other' (ibid., p. 98). This is something that we need to further explore. The claim seems to be that the body inhabits its own kind of space, while at the same time being the origination point for the perceptual space within which the things of the world appear. Are these two different kinds of space?

In fact, we need to distinguish three kinds of spatial frames of reference. The standard distinction between allocentric and egocentric spatial frames of reference names two of them. *Allocentric space* is purely objective space that can be defined in terms of latitude and longitude (the global positioning system operates in allocentric terms) or in terms of compass directions, as when we say that Copenhagen is north of Rome, for example. Once you adopt the canonical mapping of the earth, it doesn't matter where you happen to be standing, in Copenhagen, or Rome, or New York; Copenhagen is always north of Rome. *Egocentric space*, in contrast, is the perspectival space of perception and action that is defined relative to the perceiving or acting body. My computer is in front of me; the window through which I hear the church bell is to my left, and the door of my office is to my right. If I turn 180 degrees, then all of this changes. My computer is then behind me; the window is to the right and the door is to the left. This egocentric frame of reference is really a body-centred frame of reference. Kant recognized the practical importance of this egocentric, experiential spatial frame of reference:

> [. . . T]he most precise map of the heavens, if it did not, in addition to specifying the positions of the stars relative to each other, also specify the direction by reference to the position of the chart relative to my hands, would not enable me, no matter how precisely I had it in mind, to infer from a known direction, for example the north, which side of the horizon I ought to expect the sun to rise. The same thing holds of geographical and, indeed, of our most ordinary knowledge of the position of places. Such knowledge would be of no use unless we could also orientate the things thus ordered, along with the entire system of their reciprocal positions, by referring them to the sides of our body. (1992, pp. 367–68)

Simply put, even if I know that Copenhagen is to my North, if I don't know where north is relative to the way I am facing, then I don't know which way Copenhagen is. Even closer to home, so to speak, I perceive the world as organized around my body—some things to the left, some things to the right, some are up and some are down, some are in front of, and some behind me. When I reach for something, I have to reach forward or backward, to my right or to my left, at a certain angle relative to where my hand is currently positioned. Both perception and action are calibrated in egocentric space, sometimes referred to by phenomenologists as lived space. But egocentric space is still not the space of the body that Merleau-Ponty mentions.

As perceivers and agents we are embedded and embodied agents. All perception and action involve a component of bodily self-experience. I am sitting in a restaurant. I wish to begin to eat, and so I need to pick up my fork. But how can I do that? In order to pick up the fork, I need to know its position in relation to myself. That is, my perception of the fork must contain some information about me, otherwise I would not be able to act on it. On the dinner table, the perceived fork is to the left *of me*, the perceived knife is to the right *of me*, and the perceived plate and wineglass in front *of me*. This self-referencing in perception registers sub-personally in the sensorimotor system, but it also shapes my experience. Every perspectival appearance implies that the embodied perceiver is herself the experiential zero-point, the indexical 'here' in relation to which every appearing object is oriented. As an experiencing, embodied subject I am the point of reference in relation to which all of my perceptual objects are uniquely related. I am the centre around which and in relation to which (egocentric) space unfolds itself, or as Merleau-Ponty would put it, when I perceive the world, the body is simultaneously revealed as the unperceived term in the centre of the world toward which all objects turn their face.

According to Merleau-Ponty, 'the spatiality of the body must work downwards from the whole to the parts, the left hand and its position must be implied in a global bodily *design* and must originate in that design' (1962, p. 99, translation modified). He cautions, however, that this description is inadequate insofar as it remains tied to a static geometrical perspective. He suggests that we flesh this out in terms of pragmatic action: since my body is geared towards existing or possible tasks, its spatiality 'is not, like that of external objects or like that of 'spatial sensations,' a *spatiality of position*, but a *spatiality of situation*' (ibid., p. 100).[3] We should say, then, that in connection with perception and action, there is a bodily spatial frame of reference that is innate and, in its own way absolute. It is neither allocentric nor egocentric, but a frame of reference that applies to the lived body as perceiver and actor. In precise terms, this is a non-perspectival, proprioceptive frame of reference. Let's try to map this bodily space out in further detail.

The body, as Merleau-Ponty already indicated, is the origin of phenomenally experienced spatiality: 'far from my body's being for me no

more than a fragment of space, there would be no space at all for me if I had no body' (1962, p. 102). If one accepts the premise that sense perception of the world is egocentrically organized by reference to the perceiver's bodily position, the basis for that reference cannot itself be in an egocentric frame of reference without the threat of infinite regress. I could not say, for example, that my body is to my right or to my left.[4] This point is closely tied to the notion of the experiential transparency of the body (the fact that when I experience the world my experience of my body is highly attenuated), and is accurately stated by Merleau-Ponty:

> I observe external objects with my body, I handle them, examine them, walk around them, but as for my body, I do not observe it itself [in action or in the act of perception]: in order to be able to do so, I should need the use of a second body which itself would be unobservable.
> (1962, p. 91, translation modified)

Whereas I can approach or move away from any object in the world, the body itself is always here as my very perspective on the world. That is, rather than being simply another perspectivally given object, the body itself is precisely that which allows me to perceive objects perspectivally (see Sartre 1956, p. 329). In a primary sense, I am not conscious *of* my body as an intentional object. I do not perceive it; *I am it.* As a perceiver and actor, I do not have observational access to my body in perception or in action. I neither stand outside nor inside my own body—indeed, whatever inside and outside mean in this case, they depend on me being my body (see Legrand 2006).

Although I do not have observational access to my body in action, I can have non-observational proprioceptive and kinaesthetic awareness of my body in action.[5] Proprioception is the innate and intrinsic position sense that I have with respect to my limbs and overall posture. It's the 'sixth sense' that allows me to know whether my legs are crossed, or not, without looking at them. It is literally innate insofar as the proprioceptive system develops prenatally. What kind of spatial frame of reference is involved in proprioceptive awareness? It is not egocentric since proprioceptive awareness does not organize the differential spatial order of the body around a perspectival origin. For example, whereas it may be that this book is closer to me than that book over there, it is not the case that my foot is closer to me than my hand. As José Luis Bermúdez points out, there is a 'fundamental disanalogy between the bodily space of proprioception and the egocentric space of perception and action. . . . In contrast with vision, audition, and the other canonically exteroceptive modalities, there are certain spatial notions that do not seem to be applicable to somatic proprioception' (1998, pp. 152–53). Specifically he mentions distance and direction. That is, we can ask about the distance and direction of a perceived object in terms of how far away it is, and in what direction. But these spatial parameters are meaningful only in relation to a frame of reference that has a perspectival origin. This does not apply to proprioception.

Of course it is possible to read egocentric registers into the body, and to say that bodily sensation A is to the left of bodily sensation B, or that sensation A is further away from sensation B than is sensation C. Relative to a certain task (e.g. scratching an itch) my hand may be further away from my foot than from my knee, depending on posture. Someone might tell me to hold my hands out in front of me, and I might comply by extending my arms so that my hands are in front of my chest. But this is simply to adopt a certain convention or to make my chest something like a temporary origin; quite literally, one cannot put one's hands in front of one's body since they are part of the body and cannot be put in front of themselves. Being located on the front side of my body (my nose, my toes, etc.) is not equivalent to being in front of my body. Left, right, centre, and distance are spatial parameters that are completely relative in egocentric spatial perception. What is to my right may be to your left. And what is to my right now will be to my left if I turn 180 degrees. But intra-bodily, my right hand is proprioceptively just so, and always at the end of my right arm, whether my right side is located to your left, or whether I turn from north to south. If I move my left hand to touch my right shoulder, it does not become a second right hand because it happens to move to that side of my body. If sensation A is just this distance from sensation B, I cannot make them closer on the intra-bodily map even if I contort my body to make them closer objectively or pragmatically (in order to scratch one of them, for example). So intra-bodily spatiality is not egocentric.

One might think, then, that the proprioceptive frame of reference is an allocentric one. It is certainly possible to conceive of body parts being located on an allocentric map, but to the extent that allocentric means something like 'independent of the perceiver's position', it is

difficult to think of the proprioceptive mapping in those terms since it is precisely the perceiver's body that is at stake. Brian O'Shaughnessy 1995 suggests that proprioception is a system of spatial ordering that is unique in that it is framed by and applies to only the body itself. He attributes this to the immediacy of proprioception: the fact that proprioceptive awareness does not attentively mediate the perception of the body; for if it did, it would require an ordering system, a spatial frame of reference that would have to be independent of the body. Proprioception thus involves neither an allocentric nor an egocentric frame of reference, but a non-perspectival awareness of the body in an implicit spatial frame of reference.[6]

This proprioceptive frame of reference, then, is the necessary embodied basis for the egocentric frame of reference. I perceive that something is to my right or to my left only by having a proprioceptive sense of where my right is and where my left is, 'knowing' my right hand from my left hand, my right leg from my left leg. Egocentric spatial order, then, always runs back to the body of the perceiver/actor. As Merleau-Ponty tells us: 'for us to be able to conceive space, it is in the first place necessary that we should have been thrust into it by our body, and that it should have provided us with the first model of those transpositions, equivalents and identifications which make space into an objective system and allow our experience to be one of objects, opening out on an 'in itself'' (1962, p. 142). Moreover, this bodily space, in contrast to perceived space, is like 'the darkness needed in the theatre to show up the performance' (ibid., p. 100).

The Body as Experientially Transparent

Let's shine a little light into this 'darkness,' or what we might call the disappearing act of the body. We have indicated that in action, when we are engaged in some project, sensory feedback about our own body is attenuated (see Tsakiris and Haggard 2005).

> The bodily mediation most frequently escapes me: when I witness events that interest me, I am scarcely aware of the perceptual breaks which the blinking of the eye-lids imposes on the scene, and they do not figure in my memory. . . . [T]he body proper and its organs remain the bases or vehicles of my intentions and are not yet grasped as 'physiological realities.'
>
> (Merleau-Ponty 1963, p. 188, also see p. 217)

The body tries to stay out of our way so that we can get on with our task; it tends to efface itself on its way to its intentional goal. We do not normally monitor our movements in an explicitly conscious manner, although . . . we have a pre-reflective awareness of our body in very general terms. I can say whether I am running, walking, sitting, standing, and what kind of effort or posture I am putting forth. But this pre-reflective awareness is not very detailed. I can say that I am reaching to grasp a cup; but my sense of this is oriented toward the goal or intentional project that I am involved in, and not toward the specifics of my movement. I can't say very much at all about how I shape my hand in order to pick up the cup. As Sartre puts it, when I reach out to grasp something that has caught my attention, 'my hand has vanished; it is lost in the complex system of instrumentality in order that this system may exist' (1956, p. 323). He suggests that the lived body is invisibly present, precisely because it is existentially lived rather than known (ibid., p. 324). When I play ping-pong, my movements are not given as intentional objects. My limbs do not compete with the ball for my attention. If that were the case, I would be unable to play efficiently. . . . Our attention, our intentional focus, is normally on the task to be performed, the project to be accomplished, or on some worldly event that seems relevant to our action. Our attention is not on our bodily movement. Much of the action is controlled by body-schematic processes below the threshold of consciousness. Our hand shapes itself when we are grasping, and it does so automatically and without our explicit awareness. Our gait automatically adjusts itself to the terrain of the environment. This kind of automaticity, however, is not simple reflex movement. It is part of our intentional action which involves grasping something for some purpose, or choosing to stroll or to rush to our destination. Furthermore, when I execute movements, even if certain details of the control processes remain non-conscious, the movements themselves are not non-conscious, or merely mechanical, or involuntary; rather, they are part of my functioning intentionality, and are immediately and pre-reflectively felt (Henry 1975, p. 92; Merleau-Ponty 1962, p. 144).

Body image and body schema

Two concepts frequently used across a number of disciplines (both scientific and philosophical) are the concepts of *body image* and *body*

schema. Unfortunately the usage of both concepts has been rather ambiguous and confusing. In the phenomenological literature this situation has not been improved by the fact that Merleau-Ponty's term *schéma corporel* has been rendered as 'body image' in the English translation of *Phenomenology of Perception* (cf. Merleau-Ponty 1962, p. 98). We propose the following characterization. A *body image* is composed of a system of experiences, attitudes, and beliefs where the object of such intentional states is one's own body. Studies that involve body image frequently distinguish among three intentional elements:

(1) A subject's *perceptual* experience of his/her own body.
(2) A subject's *conceptual* understanding (including folk and/or scientific knowledge) of the body in general.
(3) A subject's *emotional* attitude toward his or her own body.

Conceptual and emotional aspects of body image are no doubt affected by various cultural and interpersonal factors, but in many respects their content originates in perceptual experience.

By contrast, the concept of *body schema* includes two aspects: (1) the close-to-automatic system of processes that constantly regulates posture and movement to serve intentional action; and (2) our pre-reflective and non-objectifying body-awareness. So, on the one hand, the body schema is a system of sensorimotor capacities and activations that function without the necessity of perceptual monitoring. Body-schematic processes are responsible for motor control, and involve sensorimotor capacities, abilities, and habits that enable movement and the maintenance of posture. Such processes are not perceptions, beliefs, or feelings, but sensorimotor functions that continue to operate, and in many respects operate best, when the intentional object of perception is something other than one's own body. On the other hand, however, the body schema (and this reflects Merleau-Ponty's usage of the term) also includes our pre-reflective, proprioceptive awareness of our bodily action. In either case, the emphasis is on the fact that the normal adult, in order to move around and act in the world, neither needs nor has a constant body percept that takes the body as an object. Rather, in the self-movement of most intentional activities the body-in-action tends to efface itself and to be experientially attenuated (see Gallagher 1986; Leder 1990). To the extent that one does become explicitly aware of one's own body in terms of monitoring or directing perceptual attention to limb position, movement, posture, pleasure, pain, kinaesthetic experience, and so on, such awareness constitutes aspects of a body image and presupposes the tacit contribution of the body schema.

This can be seen clearly in pathologies that involve body schematic processes. IW is a dramatic example. At the age of 19, due to illness, he lost all sense of touch and proprioception from the neck down (Cole 1995; Gallagher and Cole 1995). Shortly after the onset of his disorder, when IW tried to move a limb or his entire body, he could initiate the movement, but had no control over where the moving part ended up. If he reached for something the hands would miss or overshoot wildly, and unless he kept an eye on his hands, they started 'wandering' without his knowledge. His hands would no longer be where IW thought they were and could only be retrieved through vision. IW's lack of proprioceptive feedback has two effects. First, the normal pre-reflective proprioceptive awareness of one's bodily movement is no longer operative in IW. Second, his body schematic system, responsible for motor control, is never updated, and in effect, his body cannot gain the motor control it needs to perform action in the normal way. Subsequently IW learned to control his movements, but only through intense mental concentration and constant visual vigilance. That is, he learned to rely on a combination of *visual proprioception* and *visual perception* of limb movements, and this enabled him to move around. His awareness of his own body remains completely transformed, however. It is a reflective awareness rather than a pre-reflective one. Every single movement has to be done attentively. Even to sit in a chair without falling out of it requires constant attention. He can only initiate a standing position if he looks at his feet, and, unless he freezes in place, he can easily fall if he closes his eyes or if the lights go out. If he sneezes while walking his mental concentration is disrupted, and he could fall over. IW demonstrates exactly how much we depend on our pre-reflective, proprioceptive-kinaesthetic awareness of our bodily movement, and on body-schematic processes for the performance of action.

This body schematic aspect of embodiment constitutes what Husserl called the 'I can,' that is, the embodied capabilities for action that

correlate with the affordances of the world. As we saw in the chapter on perception, the hidden sides of objects can become present if certain *movements* are executed. Whereas the profile currently presented by the object is correlated to my present bodily position, the absent profiles are all correlated to positions that I could adopt, and this means, Husserl indicates, that they are correlated to my kinaesthetic (sensorimotor) system. I would be unable to intend the absent profiles of the object, and consequently be unable to perceive objects per se, if I were not in possession of a bodily, kinaesthetic, self-awareness in the form of an 'I can.' I 'know' my body first as a set of abilities which are not fully present to consciousness (Buytendijk 1974, p. 25)—certainly a prelinguistic and non-conceptual form of knowledge, or know-how—the limits of which become more explicitly known when things go wrong.

Imagine that you are playing tennis. Your attention is directed at the ball, which is heading towards you with high speed, as well as on the position of your opponent. Your body tightens in order to return the ball in a masterful smash, but suddenly you feel a sharp and intense pain in your chest. Your smashing opportunity is lost, and the pain is now demanding all your attention. It attracts your attention whether you want it to or not. Everything that was important a moment ago—the ball, the match, your opponent—lose significance. There is nothing that reminds us of our embodiment (our vulnerability and mortality) as much as pain. Moreover, the painful body can occasionally be experienced as alien. This is so because in pain we often lose control over the body; the 'I can' suddenly dissipates, and this disturbs the projects which define who we are (Leder 1990). Something similar is also true for various forms of illness, whether they require us to stay in bed, or observe a strict diet, or force us to visit the hospital for daily treatments.

As is frequently the case in life, it is the privation which teaches us to appreciate what we take for granted. It is when it no longer functions smoothly that we realize the importance of the body. Bernard Toussaint makes this point clearly:

> The body shows itself precisely when my body limits do not accord with the possibilities I project. . . . In such cases my body calls attention to itself as an obstacle, or as Plato would say, a prison. Thus my body becomes like an object, something alien to my intention. There arises a dichotomy between aspiration and my facticity, between project and limit. This dichotomy, I suspect, may well be the phenomenological basis for the development of the mind–body dualism.
>
> (1976, p. 176)

The lived body does not live this dualism, but when this dualism is generated—when action breaks down and our body suddenly seems to be an object in our way—we gain some phenomenological access to what generally goes unnoticed—the smooth functioning of our body in perception and action as the constant and pervasive support system for our cognitive life.

Embodiment and Social Cognition

We will have more to say about the relation between embodiment and intersubjectivity. . . but let us end this chapter with at least some indications of the link. It should be obvious that my bodily self-apprehension and the way I live my body can be influenced by my social interaction, and by the way my body is perceived and apprehended by others—just think of categories like gender and race. But perhaps even more basically, social interaction is as such an embodied practice.

To exist embodied is to exist in such a way that one exists under the gaze of the other, accessible for the other. My bodily behaviour always has a public side to it. Thus the standard question which is posed as 'the problem of other minds'—'How do I find an access to the other'—is mistaken. It signals that I am enclosed in my own interiority, and that I then have to employ methods to reach the other who is outside. But this way of framing the problem fails to recognize the nature of embodiment.

Bodily behaviour, expression, and action are essential to (and not merely contingent vehicles of) some basic forms of consciousness. Mental states do not simply serve to explain behaviour; rather, some mental states are directly apprehended in the bodily expressions of people whose mental states they are. As Hobson has recently put it: 'We perceive bodies and bodily expressions, but we do so in such a way that we perceive and react to the mental life that those physical forms express' (2002, p. 248; cf. 1993, p. 184).

When presented with behaviour, it is not as if we are faced with mere bodily processes that can then be interpreted any way one likes. Rather, it is more like being confronted with

a language. Even a foreign and incomprehensible language is perceived as meaningful and not simply as physical noise. When you see somebody use a hammer, or feed a child, or clean a table you don't have a problem understanding what is going on. You don't necessarily understand every aspect of the action, but it is immediately given as a meaningful action (in a shared world). It is not as if you are first confronted with a perceived exterior, and then have to infer the existence of an interior mental space. In the face-to-face encounter, we are neither confronted with a mere body, nor with a hidden psyche, but with a unified whole. When I see another's face, I *see* it as friendly or angry, etc., that is, the very face expresses these emotions. This does not rule out that some mental states are covert, of course, but not all mental states can lack an essential link to behaviour if intersubjectivity is at all to get off the ground.

To take embodiment seriously is to contest a Cartesian view of the mind in more than one way. Embodiment entails birth and death. To be born is not to be one's own foundation, but to be situated in both *nature* and *culture.* It is to possess a physiology that one did not choose. It is to find oneself in a historical and sociological context that one did not establish (see Merleau-Ponty 1962, p. 347). Birth is essentially an intersubjective phenomenon, not only in the obvious sense, because I was born by somebody, but because this very event only has meaning for me through others. My awareness of my birth, of my commencement, and of my mortality, is intersubjectively mediated; it is not something I can intuit or remember on my own. I do not witness my coming into being, but I always already find myself alive (Merleau-Ponty 1962, p. 215; Ricoeur 1966, pp. 433, 438, 441). Ultimately, the issues of birth and death enlarge the scope of the investigation. They call attention to the role of historicity, generativity, and sexuality.[7] Indeed, rather than being simply a biological given, embodiment is also a category of sociocultural analysis. What this means, however, is that to gain a more comprehensive understanding of the embodied mind, one needs to take a much wider scope, and the first step in developing this expanded concept of the mind is to consider the complexities of the circumstances in which more than one body is involved and where there is intersubjective interaction. Before we look at interaction, however, it will be helpful to look at action itself. Intersubjectivity is not found simply in the proximity of two or more passive subjects, but is also an encounter between agents.

FURTHER READING

Bermúdez, José Luis, Anthony Marcel, and Naomi Eilan, eds., *The Body and the Self* (Cambridge, MA: MIT Press, 1995).

Clark, Andy. *Being There: Putting Brain, Body, and World Together Again* (Cambridge, MA: MIT Press, 1997).

Gallagher, Shaun. *How the Body Shapes the Mind* (Oxford: Oxford University Press/Clarendon Press, 2005).

Henry, Michel. *Philosophy and Phenomenology of the Body,* G. Etzkorn, trans., (The Hague: Martinus Nijhoff, 1975).

Husserl, Edmund. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy*, R. Rojcewicz and A. Schuwer, trans., (Dordrecht: Kluwer Academic Publishers, 1989), Second Book.

Leder, Drew. *The Absent Body* (Chicago: Chicago University Press, 1990).

Sheets-Johnstone, Maxine. *The Primacy of Movement* (Amsterdam: John Benjamins, 1999).

Varela, Francisco, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge, MA: MIT Press, 1991).

Waldenfels, Bernard. *Das leibliche Selbst. Vorlesungen zur Phänomenologie des Leibes* (Frankfurt am Main: Suhrkamp, 2001).

Wider, Kathleen V. *The Bodily Nature of Consciousness: Sartre and Contemporary Philosophy of Mind* (Ithaca, NY: Cornell University Press, 1997).

NOTES

1. For an illuminating discussion of Sartre's analysis of the body, cf. Cabestan 1996.
2. This pre-reflective sense of our own embodiment contributes to our ability to identify our body in an objective fashion. Subjects who view videos of moving figures wearing point-light displays in the dark (the bodies are marked by lights positioned at joints so that when walking their gait is clearly visible) are better at identifying themselves than they are at recognizing friends and colleagues. The puzzle is how they are able to do so, since what they see is how their gait looks 'from the outside.' And this cannot be

something they are perceptually well acquainted with, since people obviously see the gaits of friends and colleagues more than they see their own (Gibbs 2006, p. 51). But one suggestion is that their pre-reflective proprioceptive sense of their own embodiment cross-modally informs their perception of the visual Gestalt of their gait. For more on the pre-reflective awareness of one's body, see Legrand 2006.

3. In his lectures at the Collège de France in the mid-1950s, for example, Merleau-Ponty describes perception in this context in the following terms. When I perceive an object, I am aware of it in terms of my implicit motor possibilities: 'The thing appears to me as a function of the [actual and potential] movements of my body. . . . My body is the absolute 'here.' All the places of space proceed from it. . . . The Absolute in the relative is what my body brings to me' (2003, pp. 74–75).
4. Cases of out-of-body experiences (OBEs) or autoscopy (AS), where an awake person apparently sees his body from a position outside of it, are special cases where, to get the phenomenology correct, one needs to distinguish between the lived or perceiving body and the objective or perceived body. Blanke et al. 2004 suggest that OBE and AS involve a failure to integrate proprioceptive, tactile, and visual information with respect to one's own body, and a vestibular dysfunction that leads to an additional disintegration between personal (vestibular) space and extrapersonal (visual) space.
5. Phenomenologists take pre-reflective body-awareness to be a question of how (embodied) consciousness is given to itself not as an *object*, but as a *subject*. Whereas Bermúdez has recently claimed that 'somatic proprioception is a form of perception' that takes 'the embodied self as its object' (1998, p. 132), the phenomenologists would argue that *primary* body-awareness is not a type of object-consciousness; it is *not* a perception of the body as an object at all (cf. Gallagher 2003b; Zahavi 2002).
6. The proprioceptive register is thus not independent of the subject's experience. The grid of a global positioning system might map out my body as I lay in the sun at Cocoa Beach, but this is not the system I use when I need to scratch my foot; nor do I have to figure out whether my foot is to the East or West of my hand, even if in some languages I would have to figure out whether the foot was on my South leg. '[S]peakers of Guugu Yimithirr (Australia) use only the last kind of description; . . . even to describe the location of an object on a body part—a Guugu Yimithirr speaker would say 'There's an ant on your south leg'' (Majid et al. 2004, pp. 108–9). Majid et al., however, ignore the question of how this knowledge is possible. Compass directions can only be dead reckoned from the 'here' of my lived body in relation to some landmark. Directions are always directions away from *me*, where my 'here' defines the 'first co-ordinates.' Phenomenologically, I triangulate north, pointing to it from here, with some implicit or explicit reference point of which I know the relative location. How do I know whether the foot I want to scratch is on my south leg or my north leg? I have to determine whether my right leg is to the north or south of my left leg, and to do that I first have to know whether the northerly direction is to my right or to my left (see Gallagher 2006).
7. Heidegger, who is not exactly known as a philosopher of the body, chose a neuter, *'das Dasein,'* as the central term for human existence. And as Heidegger points out in the lecture course *Metaphysische Anfangsgründe der Logik im Ausgang von Leibniz* from 1928, the neutrality of *Dasein* entails an asexuality (eine *Geschlechtslosigkeit*) (Heidegger 1978, p. 172). Subsequent thinkers have questioned the validity of this move, and have argued that the basic structures of our embodiment would not remain the same if we were asexual creatures.

4 Perception

Perception is a central aspect of the mind. We gather information about things in the world around us by seeing them, hearing them, touching them, and so on. Much of the time our consciousness is dominated by perceptual experience: the experience of seeing, hearing, touching, and so on. Empiricists in early modern philosophy such as John Locke, George Berkeley, and David Hume held that perception was the basis of all knowledge.

At the same time, perception itself is ill-understood. What are the objects of perception? And what sort of state or process is perceptual experience itself? It is natural to hold that the objects of perception are objects in the world around us, like tables and chairs, at least in ordinary cases of perception. And it is natural to hold that perception itself is a relation to those objects. Views of this sort are sometimes called *naïve realism* about perception.

The 20th-century discussion of perception has often started by thinking about cases where perception goes wrong. These cases include *illusion*, such as the case where a stick in water looks bent where it is straight, or two lines look to be different lengths when they are the same length. In these cases, there is an external object one perceives but one perceives it as being a way that is not. They also include *hallucination*, such as a case where one seems to see a pink elephant in front of one, but there is nothing there at all. In these cases, there is no external object one perceives. The object of perception is merely apparent. Thinking about cases of illusion and hallucination has led many (but not all) philosophers away from naïve realism toward other views of perception.

In an excerpt from his 1940 book *The Foundations of Knowledge*, the British philosopher A. J. Ayer (chapter 52) gives a succinct version of the *argument from illusion.* When we perceive the stick in water, the object we perceive is *bent.* But the actual stick in the water is not bent but straight. So we do not perceive the actual stick. Instead we perceive a different object, a *sense-datum*, which is genuinely bent. Similarly, when we perceive the two lines, we perceive sense-data which are of the same length. And when we hallucinate the elephant, we perceive an elephantine sense-datum. Furthermore, Ayer argues, ordinary cases of perception are continuous with cases of illusion and hallucination and should be treated in the same way. If so, then even in ordinary perception, we directly perceive sense-data, and only indirectly perceive objects in the external world.

The sense-datum view of perception was held by founders of analytic philosophy such as Bertrand Russell and G. E. Moore, and was popular for much of the 20th century. At the same time it raises many puzzles. Where are sense-data? (In the world? In the mind? Somewhere else?) What sort of thing are they? (Physical objects? Mental objects? Something else?) What properties do they have? (Colors and shapes? Being a stick or an elephant?) Much of 20th-century philosophy of perception has involved finding paths away from the sense-datum view.

The sense-datum theory and many other approaches to perception give a central role to the *sensation* as the basic unit of perception. We have sensations of red or green, of hot or of cold. In chapter 54, which is itself the first chapter of his 1945 book *The Phenomenology of Perception*, the French phenomenologist Maurice Merleau-Ponty argues that the idea of a sensation is misconceived. There is no such thing as a sensation of red or of cold taken alone. These things are always perceived against a background of more going on, as part of a whole perceptual field. This leads Merleau-Ponty toward an approach to perception that gives a central role to the body.

Perhaps the most prominent critique both of the sense-datum view and of the argument for illusion was given by the Oxford philosopher J. L. Austin in his lectures on *Sense and Sensibilia* (given in the 1950s, and published after his death in 1962). Austin was a leader of the school of ordinary-language philosophy, who held that philosophical reasoning must respect the way that ordinary language is used. In the material excerpted here (chapter 53), Austin argues that Ayer's argument do violence to ordinary language at a number of points, for example in talking being 'deceived by their senses' or being subject to 'illusions.' He suggests that we should not conflate the distinct categories of illusion or delusion. And he argues that in the case of the stick, it would be absurd to say (as Ayer's sense-datum theory does) that we perceive an immaterial thing. What we see is a stick partly immersed in water.

The British philosopher G. E. M. Anscombe (chapter 55) argues that sensation and perception possess the key features of *intentionality*: that is, they are directed at the world. Anscombe explores the intentionality involved when we talking of thinking of someone or worshipping someone: for example, we can worship Zeus even though Zeus does not exist. In this case, we say Zeus is the intentional object of worshipping. Then she argues that talk of sense-perception exhibits the same features: for example, we can hear a ringing in our ears even if there is no ringing in our ears. She argues that like thinking and worshiping, sense perception has intentional objects. She uses this to draw out problems both for the Ayer's sense-datum view and Austin's ordinary-language view.

Anscombe's observations about the intentionality of sensation were a predecessor of the view that perceptual experience is itself a sort of intentional or representational state: a view known as *intentionalism* or *representationalism.* We have already explored representationalism as a theory of consciousness in chapter 27 by Michael Tye in part 2. Tye's chapter serves equally as a representative of intentionalism about perception. He sets out motivations for the view, based especially in the transparency of perception, draws out commitments of the view, and answers a number of objections.

The leading rival of intentionalism in contemporary philosophy of mind is not the sense-datum view but naïve realism, where at least ordinary cases of perceptual experience involve a direct relation to external objects. Naïve realism is often spelled out in a form involving *disjunctivism*, on which successful perception and cases of illusion and hallucination are fundamentally distinct from each other and involve different sorts of perceptual experience. In chapter 56, M. G. F. Martin spells out many of the commitments of a disjunctivist naïve realist view and contrasts it with other views of illusion and hallucination.

Much of the recent philosophy of perception has been driven by the science of perception. On central body of scientific work concerns *change blindness* and *inattentional blindness*, where it has been shown that subjects frequently do not notice objects and events that they are not paying attention to. A number of philosophers and scientists have argued that in these cases, we do not have perceptual experiences of the objects at all. We are only conscious of a limited number of objects that we are attending to. Some of these philosophers have argued that we suffer from a *grand illusion* saying that we have detailed perception of many objects, when in fact we have sparse perception of a few.

In chapter 57, Alva Noë takes on the grand illusion view and argues that it is mistaken. He argues that we do in fact perceive a wide array of objects in our environment. He argues that we should reject a 'snapshot' view of perception and recognize that what we perceive goes well beyond what is immediately given to us. Instead we should endorse an *enactive* or *sensorimotor* view of perception where perception depends on

our relation to action. For example, we perceive the part of a cat that is behind a fence because we know that if we move our head, it will come into view. As a result, the links between perception and action are deeper than usually thought.

Perception seems to involve awareness of both objects (such as a stick) and properties (such as its bentness). The chapters so far have focused mostly on objects, but perceptual awareness of properties is equally important. Susanna Siegel (chapter 58) asks: which properties are represented in perception? It is often held (especially by sense-datum theorists, but also by many representationalists and naïve realists) that we are perceptually aware of just a small range of properties: in the case of vision, perhaps just spatial properties (size, shape, location) and color and illumination. Against this view, Siegel argues that we can be aware of a wide range of properties, such as being a house, being a pine tree, or being a specific person. As a result the contents of perception are richer than often thought.

FURTHER READING

Bertrand Russell 1912 gives a classic statement of the sense-datum view. Another famous critique of the view is Wilfred Sellars' discussion of the "Myth of the Given" in the opening sections of "Empiricism and the Philosophy of Mind" (1958). Jackson 1978 and Robinson 1994 defend sense-datum views. Representationalist views are explored at length by Dretske 1995 and Tye 1995. Disjunctivist views were introduced by Hinton 1968 and developed further by McDowell 1995, Fish 2009, Brewer 2011, and many others. Schellenberg 2018 develops a view of perception with both representationalist and disjunctivist elements. Noë's and Siegel's distinctive views of perception are developed at much more length in Noë 2004 and Siegel 2011. Other important recent book-length treatments of perception include Brogaard 2018. O'Callaghan 2020 and Orlandi 2014. Many interesting articles on perceptual experience are collected in Crane 1992 and Gendler and Hawthorne 2007.

Brewer, B. *Perception and its objects* (Oxford: Oxford University Press, 2011).

Brogaard, B. *Seeing and Saying: The Language of Perception and the Representational View of Experience* (New York: Oxford University Press, 2018).

Crane, T. *The Contents of Experience: Essays on Perception* (Cambridge, MA: Cambridge University Press, 1992).

Dretske, F. *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1994).

Fish, W. *Perception, Hallucination, and Illusion* (Oxford: Oxford University Press, 2009).

Gendler, T., and Hawthorne, J. *Perceptual Experience* (Oxford: Oxford University Press, 2007).

Hinton, J. M., "Visual experiences," *Mind* 76 (1967): pp. 217–27.

Jackson, F. *Perception: A Representative Theory* (Cambridge, MA: Cambridge University Press, 1977).

McDowell, J. *Mind and World* (Cambridge, MA: Harvard University Press, 1994).

O'Callaghan, C. *A Multisensory Philosophy of Perception* (New York: Oxford University Press, 2020).

Orlandi, N. *The Innocent Eye: Why Perception Is not a Cognitive Process* (New York Oxford University Press, 2014).

Pautz, A. *Perception* (London: Routledge, 2017).

Robinson, H. *Perception* (London: Routledge, 1994).

Russell, B. *The Problems of Philosophy* (Home University Library, 1912).

Schellenberg, S. *The Unity of Perception* (Oxford, UK: Oxford University Press, 2018).

Sellars, W. *Empiricism and the philosophy of mind* (1958). Reprinted as *Empiricism and the Philosophy of Mind* (Cambridge, MA: Harvard University Press, 1997).

Siegel, S. *The Contents of Visual Experience* (Oxford, UK: Oxford University Press, 2011).

Tye, M. *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995).

The Argument from Illusion

A. J. Ayer

Exposition of the Argument

It does not normally occur to us that there is any need for us to justify our belief m the existence of material things. At the present moment, for example, I have no doubt whatsoever that I really am perceiving the familiar objects, the chairs and table, the pictures and books and flowers with which my room is furnished, and I am therefore satisfied that they exist. I recognize indeed that people are sometimes deceived by their senses, but this does not lead me to suspect that my own sense-perceptions cannot in general be trusted, or even that they may be deceiving me now. And this is not, I believe, an exceptional attitude. I believe that, in practice, most people agree with John Locke that 'the certainty of things existing *in rerum natura*, when we have the testimony of our senses for it, is not only as great as our frame can attain to, but as our condition needs.[1]

When, however, one turns to the writings of those philosophers who have recently concerned themselves with the subject of perception, one may begin to wonder whether this matter is quite so simple. It is true that they do, in general, allow that our belief in the existence of material things is well founded; some of them, indeed, would say that there were occasions on which we knew for certain the truth of such propositions as 'this is a cigarette' or 'this is a pen.' But even so they are not, for the most part, prepared to admit that such objects as pens or cigarettes are ever directly perceived. What, in their opinion, we directly perceive is always an object of a different kind from these, one to which it is now customary to give the name of 'sense-datum.' These sense-data are said to have the 'presentative function'[2] of making us conscious of material things. But how they perform this function, and what is their relation to the material things which they present, are questions about which there is much dispute. There is dispute also about the properties of sense-data, apart from their relationship to material things: whether, for example, they are each of them private to a single observer, whether they can appear to have qualities that they do not really have, or have qualities that they do not appear to have, whether they are in any sense 'within' the percipient's mind or brain. I shall show later on that these are not empirical questions. They are to be settled by making it clear how the term 'sense-datum' is intended to be used. But first I must explain why it is thought necessary to introduce such a term at all. Why may we not say that we are directly aware of material things?

The answer is provided by what is known as the argument from illusion. This argument, as it is ordinarily stated, is based on the fact that material things may present different appearances to different observers, or to the same observer in different conditions, and that the character of these appearances is to some extent causally determined by the state of the conditions and the observer. For instance, it is remarked that a coin which looks circular from one point of view may look elliptical from another, or that a stick which normally appears straight looks bent when it is seen in water, or that to people who take drugs such as mescal, things appear to change their colours. The familiar cases of mirror images, and double vision, and complete hallucinations, such as the mirage, provide further examples. Nor is this a peculiarity of visual appearances. The same thing occurs in the domains of the other senses, including the sense of touch. It may be pointed out, for example, that the taste that a thing appears to have may vary with the condition of the palate, or that a liquid will seem to have a different temperature according as the hand that is feeling it is itself hot or cold; or that a coin seems larger when it is placed on the tongue than when it is held in the palm of the hand, or, to take a case of complete hallucination, that people who have had limbs amputated may still continue to feel pain in them.

Let us now consider one of these examples, say, that of the stick which is refracted in water, and see what is to be inferred. For the present it must be assumed that the stick does not really change its shape when it is placed in water. I shall discuss the meaning and validity of this assumption later on. Then it follows that at

Excerpted from *The Foundations Of Empirical Knowledge* (Macmillan, 1940). Reprinted by permission of the publisher.

least one of the visual appearances of the stick is delusive, for it cannot be both crooked and straight. Nevertheless, even in the case where what we see is not the real quality of a material thing, it is supposed that we are still seeing something, and that it is convenient to give this a name. And it is for this purpose that philosophers have recourse to the term 'sense-datum.' By using it they are able to give what seems to them a satisfactory answer to the question, 'What is the object of which we are directly aware, in perception, if it is not part of any material thing?' Thus, when a man sees a mirage in the desert, he is not thereby perceiving any material thing, for the oasis which he thinks he is perceiving does not exist. At the same time, it is argued, his experience is not an experience of nothing, it has a definite content. Accordingly, it is said that he is experiencing sense-data, which are similar in character to what he would be experiencing if he were seeing a real oasis, but are delusive in the sense that the material thing which they appear to present is not actually there. Or again, when I look at myself in the glass, my body appears to be some distance behind the glass; but other observations indicate that it is in front of it. Since it is impossible for my body to be in both these places at once, these perceptions cannot all be veridical. I believe, in fact, that the ones that are delusive are those in which my body appears to be behind the glass. But can it be denied that when one looks at oneself in the glass one is seeing something? And if, in this case, there really is no such material thing as my body in the place where it appears to be, what is it that I am seeing? Once again the answer we are invited to give is that it is a sense-datum. And the same conclusion may be reached by taking any other of my examples.

If anything is established by this, it can be only that there are some cases in which the character of our perceptions makes it necessary for us to say that what we are directly experiencing is not a material thing but a sense-datum. It has not been shown that this is so in all cases. It has not been denied, but rather assumed, that there are some perceptions that do present material things to us as they really are; and in their case there seems at first sight to be no ground for saying that we directly experience sense-data rather than material things. But, as I have already remarked, there is general agreement among the philosophers who make use of the term 'sense-datum,' or some equivalent term, that what we immediately experience is always a sense-datum and never a material thing. And for this they give further arguments which I shall now examine.

In the first place it is pointed out that there is no intrinsic difference in kind between those of our perceptions that are veridical in their presentation of material things and those that are delusive.[3] When I look at a straight stick, which is refracted in water and so appears crooked, my experience is qualitatively the same as if I were looking at a stick that really was crooked. When, as the result of my putting on green spectacles, the white walls of my room appear to me to be green, my experience is qualitatively the same as if I were perceiving walls that really were green. When people whose legs have been amputated continue to feel pressure upon them, their experience is qualitatively the same as if pressure really were being exerted upon their legs. But, it is argued, if, when our perceptions were delusive, we were always perceiving something of a different kind from what we perceived when they were veridical, we should expect our experience to be qualitatively different in the two cases. We should expect to be able to tell from the intrinsic character of a perception whether it was a perception of a sense-datum or of a material thing. But this is not possible, as the examples that I have given have shown. In some cases there is indeed a distinction with respect to the beliefs to which the experiences give rise, as can be illustrated by my original example. For when, in normal conditions, we have the experience of seeing a straight stick, we believe that there really is a straight stick there, but when the stick appears crooked, through being refracted in water, we do not believe that it really is crooked, we do not regard the fact that it looks crooked in water as evidence against its being really straight. It must, however, be remarked that this difference in the beliefs which accompany our perceptions is not grounded in the nature of the perceptions themselves, but depends upon our past experience. We do not believe that the stick which appears crooked when it stands in water really is crooked because we know from past experience that in normal conditions it looks straight. But a child who had not learned that refraction was a means of distortion would naturally believe that the stick really was crooked as he saw it. The fact, therefore, that there is this distinction between the beliefs that accompany veridical and delusive perceptions does not justify the view that these are perceptions of generically different objects, especially as the distinction by no means applies to all cases. For it sometimes

happens that a delusive experience is not only qualitatively indistinguishable from one that is veridical but is also itself believed to be veridical, as in the example of the mirage; and, conversely, there are cases in which experiences that are actually veridical are believed to be delusive, as when we see something so strange or unexpected that we say to ourselves that we must be dreaming. The fact is that from the character of a perception considered by itself, that is, apart from its relation to further sense-experience, it is not possible to tell whether it is veridical or delusive. But whether we are entitled to infer from this that what we immediately experience is always a sense-datum remains still to be seen.

Another fact which is supposed to show that even in the case of veridical perceptions we are not directly aware of material things is that veridical and delusive perceptions may form a continuous series, both with respect to their qualities and with respect to the conditions in which they are obtained.[4] Thus, if I gradually approach an object from a distance I may begin by having a series of perceptions which are delusive in the sense that the object appears to be smaller than it really is. Let us assume that this series terminates in a veridical perception. Then the difference in quality between this perception and its immediate predecessor will be of the same order as the difference between any two delusive perceptions that are next to one another in the series; and, on the assumption that I am walking at a uniform pace, the same will be true of the difference in the conditions on which the generation of the series depends. A similar example would be that of the continuous alteration in the apparent colour of an object which was seen in a gradually changing light. Here again the relation between a veridical perception and the delusive perception that comes next to it in the series is the same as that which obtains between neighbouring delusive perceptions, both with respect to the difference in quality and with respect to the change in the conditions, and these are differences of degree and not of kind. But this, it is argued, is not what we should expect if the veridical perception were a perception of an object of a different sort, a material thing as opposed to a sense-datum. Does not the fact that veridical and delusive perceptions shade into one another in the way that is indicated by these examples show that the objects that are perceived in either case are generically the same? And from this it would follow, if it was acknowledged that the delusive perceptions were perceptions of sense-data, that what we directly experienced was always a sense-datum and never a material thing.

The final argument that has to be considered m this context is based upon the fact that all our perceptions, whether veridical or delusive, are to some extent causally dependent both upon external conditions, such as the character of the light, and upon our own physiological and psychological states. In the case of perceptions that we take to be delusive this is a fact that we habitually recognize. We say, for example, that the stick looks crooked because it is seen in water; that the white walls appear green to me because I am wearing green spectacles; that the water feels cool because my hand is hot; that the murderer sees the ghost of his victim because of his bad conscience or because he has been taking drugs. In the case of perceptions that we take to be veridical we are apt not to notice such causal dependencies, since as a rule it is only the occurrence of the unexpected or the abnormal that induces us to look for a cause. But in this matter also there is no essential difference between veridical and delusive perceptions. When, for example, I look at the piece of paper on which I am writing, I may claim that I am seeing it as it really is. But I must admit that in order that I should have this experience it is not sufficient that there should actually be such a piece of paper there. Many other factors are necessary, such as the condition of the light, the distance at which I am from the paper, the nature of the background, the state of my nervous system and my eyes. A proof that they are necessary is that if I vary them I find that I have altered the character of my perception. Thus, if I screw up my eyes I see two pieces of paper instead of one; if I grow dizzy the appearance of the paper becomes blurred; if I alter my position sufficiently it appears to have a different shape and size; if the light is extinguished, or another object is interposed, I cease to see it altogether. On the other hand, the converse does not hold. If the paper is removed I shall cease to see it; but the state of the light or of my nervous system or any other of the factors that were relevant to the occurrence of my perception may still remain the same. From this it may be inferred that the relation between my perception and these accompanying conditions is such that, while they are not causally dependent upon it, it is causally dependent upon them. And the same

would apply to any other instance of a veridical perception that one cared to choose.

This point being established, the argument proceeds as follows. It is held to be characteristic of material things that their existence and their essential properties are independent of any particular observer. For they are supposed to continue the same, whether they are observed by one person or another, or not observed at all. But this, it is argued, has been shown not to be true of the objects we immediately experience. And so the conclusion is reached that what we immediately experience is in no case a material thing. According to this way of reasoning, if some perceptions are rightly held to be veridical, and others delusive, it is because of the different relations in which their objects stand to material things, and it is a philosophical problem to discover what these relations are. We may be allowed to have indirect knowledge of the properties of material things. But this knowledge, it is held, must be obtained through the medium of sense-data, since they are the only objects of which, in sense-perception, we are immediately aware.

NOTES

1. *An Essay concerning Human Understanding*, Book IV, chap. 2, section VIII.
2. Cf H. H. Price, *Perception*, p. 104.
3. Cf H. H. Price, *Perception*, p. 31.
4. Cf Price, op. cit., p. 32.

Sense and Sensibilia

J. L. Austin

I

In these lectures I am going to discuss some current doctrines (perhaps, by now, not so current as they once were) about sense-perception. We shall not, I fear, get so far as to decide about the truth or falsity of these doctrines; but in fact that is a question that really *can't* be decided, since it turns out that they all bite off more than they can chew. I shall take as chief stalking-horse in the discussion Professor A. J. Ayer's *The Foundations of Empirical Knowledge*;[1] but I shall mention also Professor H. H. Price's *Perception*,[2] and, later on, G. J. Warnock's book on Berkeley.[3] I find in these texts a good deal to criticize, but I choose them for their merits and not for their deficiencies; they seem to me to provide the best available expositions of the approved reasons for holding theories which are at least as old as Heraclitus—more full, coherent, and terminologically exact than you find, for example, in Descartes or Berkeley. No doubt the authors of these books no longer hold the theories expounded in them, or at any rate wouldn't now expound them in just the same form. But at least they did hold them not very long ago; and of course very numerous great philosophers have held these theories, and have propounded other doctrines resulting from them. The authors I have chosen to discuss may differ from each other over certain refinements, which we shall eventually take note of—they appear to differ, for example, as to whether their central distinction is between two 'languages' or between two classes of entities—but I believe that they agree with each other, and with their predecessors, in all their major (and mostly unnoticed) assumptions.

Excerpted from *Sense and Sensibilia* (Oxford University Press, 1962).

Ideally, I suppose, a discussion of this sort ought to begin with the very earliest texts; but in this case that course is ruled out by their no longer being extant. The doctrines we shall be discussing—unlike, for example, doctrines about 'universals'—were already quite ancient in Plato's time.

The general doctrine, generally stated, goes like this: we never see or otherwise perceive (or 'sense'), or anyhow we never *directly* perceive or sense, material objects (or material things), but only sense-data (or our own ideas, impressions, sensa, sense-perceptions, percepts, etc.).

One might well want to ask how seriously this doctrine is intended, just how strictly and literally the philosophers who propound it mean their words to be taken. But I think we had better not worry about this question for the present. It is, as a matter of fact, not at all easy to answer, for, strange though the doctrine looks, we are sometimes told to take it easy—really it's just what we've all believed all along. (There's the bit where you say it and the bit where you take it back.) In any case it is clear that the doctrine is thought *worth stating*, and equally there is no doubt that people find it disturbing; so at least we can begin with the assurance that it deserves serious attention.

My general opinion about this doctrine is that it is a typically *scholastic* view, attributable, first, to an obsession with a few particular words, the uses of which are over-simplified, not really understood or carefully studied or correctly described; and second, to an obsession with a few (and nearly always the same) half-studied 'facts.' (I say 'scholastic,' but I might just as well have said 'philosophical'; over-simplification, schematization, and constant obsessive repetition of the same small range of jejune 'examples' are not only not peculiar to this case, but far too common to be dismissed as an occasional weakness of philosophers.) The fact is, as I shall try to make clear, that our ordinary words are much subtler in their uses, and mark many more distinctions, than philosophers have realized; and that the facts of perception, as discovered by, for instance, psychologists but also as noted by common mortals, are much more diverse and complicated than has been allowed for. It is essential, here as elsewhere, to abandon old habits of *Gleichschaltung,* the deeply ingrained worship of tidy-looking dichotomies.

I am *not,* then—and this is a point to be clear about from the beginning—going to maintain that we ought to be 'realists,' to embrace, that is, the doctrine that we *do* perceive material things (or objects). This doctrine would be no less scholastic and erroneous than its antithesis. The question, do we perceive material things or sense-data, no doubt looks very simple—*too* simple—but is entirely misleading (cp. Thales' similarly vast and over-simple question, what the world is made of). One of the most important points to grasp is that these two terms, 'sense-data' and 'material things,' live by taking in each other's washing—what is spurious is not one term of the pair, but the antithesis itself.[4] There is no *one* kind of thing that we 'perceive' but many *different* kinds, the number being reducible if at all by scientific investigation and not by philosophy: pens are in many ways though not in all ways unlike rainbows, which are in many ways though not in all ways unlike after-images, which in turn are in many ways but not in all ways unlike pictures on the cinema-screen—and so on, without assignable limit. So we are *not* to look for an answer to the question, what kind of thing we perceive. What we have above all to do is, negatively, to rid ourselves of such illusions as 'the argument from illusion'—an 'argument' which those (e.g. Berkeley, Hume, Russell, Ayer) who have been most adept at working it, most fully masters of a certain special, happy style of blinkering philosophical English, have all themselves felt to be somehow spurious. There is no simple way of doing this—partly because, as we shall see, there is no simple 'argument.' It is a matter of unpicking, one by one, a mass of seductive (mainly verbal) fallacies, of exposing a wide variety of concealed motives—an Operation which leaves us, in a sense, just where we began.

In a sense—but actually we may hope to learn something positive in the way of a technique for dissolving philosophical worries (*some* kinds of philosophical worry, not the whole of philosophy); and also something about the meanings of some English words ('reality,' 'seems,' 'looks,' etc.) which, besides being philosophically very slippery, are in their own right interesting. Besides, there is nothing so plain boring as the constant repetition of assertions that are not true, and sometimes not even faintly sensible; if we can reduce this a bit, it will be all to the good.

II

Let us have a look, then, at the very beginning of Ayer's *Foundations*—the bottom, one might

perhaps call it, of the garden path. In these paragraphs[5] we already seem to see the plain man, here under the implausible aspect of Ayer himself, dribbling briskly into position in front of his own goal, and squaring up to encompass his own destruction.

It does not normally occur to us that there is any need for us to justify our belief in the existence of material things. At the present moment, for example, I have no doubt whatsoever that I really am perceiving the familiar objects, the chairs and table, the pictures and books and flowers with which my room is furnished; and I am therefore satisfied that they exist. I recognize indeed that people are sometimes deceived by their senses, but this does not lead me to suspect that my own sense-perceptions cannot in general be trusted, or even that they may be deceiving me now. And this is not, I believe, an exceptional attitude. I believe that, in practice, most people agree with John Locke that 'the certainty of things existing *in rerum natura,* when we have the testimony of our senses for it, is not only as great as our frame can attain to, but as our condition needs.'

When, however, one turns to the writings of those philosophers who have recently concerned themselves with the subject of perception, one may begin to wonder whether this matter is quite so simple. It is true that they do, in general, allow that our belief in the existence of material things is well founded; some of them, indeed, would say that there were occasions on which we knew for certain the truth of such propositions as 'this is a cigarette' or 'this is a pen.' But even so they are not, for the most part, prepared to admit that such objects as pens or cigarettes are ever directly perceived. What, in their opinion, we directly perceive is always an object of a different kind from these; one to which it is now customary to give the name of 'sense-datum.'

Now in this passage some sort of contrast is drawn between what we (or the ordinary man) believe (or believes), and what philosophers, at least 'for the most part,' believe or are 'prepared to admit.' We must look at both sides of this contrast, and with particular care at what is assumed in, and implied by, what is actually said. The ordinary man's side, then, first.

1. It is clearly implied, first of all, that the ordinary man believes that he perceives material things. Now this, at least if it is taken to mean that he would *say* that he perceives material things, is surely wrong straight off; for 'material thing' is not an expression which the ordinary man would use—nor, probably, is 'perceive.' Presumably, though, the expression 'material thing' is here put forward, not as what the ordinary man would *say,* but as designating in a general way the *class* of things of which the ordinary man both believes and from time to time says that he perceives particular instances. But then we have to ask, of course, what this class comprises. We are given, as examples, 'familiar objects'—chairs, tables, pictures, books, flowers, pens, cigarettes; the expression 'material thing' is not here (or anywhere else in Ayer's text) further defined.[6] But *does* the ordinary man believe that what he perceives is (always) something like furniture, or like these other 'familiar objects'—moderate-sized specimens of dry goods? We may think, for instance, of people, people's voices, rivers, mountains, flames, rainbows, shadows, pictures on the screen at the cinema, pictures in books or hung on walls, vapours, gases—all of which people say that they see or (in some cases) hear or smell, i.e. 'perceive.' Are these all 'material things'? If not, exactly which are not, and exactly why? No answer is vouchsafed. The trouble is that the expression 'material thing' is functioning *already*, from the very beginning, simply as a foil for 'sense-datum'; it is not here given, and is never given, any other role to play, and apart from this consideration it would surely never have occurred to anybody to try to represent as some single *kind of things* the things which the ordinary man says that he 'perceives.'
2. Further, it seems to be also implied (*a*) that when the ordinary man believes that he is not perceiving material things, he believes he is being deceived by his senses; and (*b*) that when he believes he is being deceived by his senses, he believes that he is not perceiving material things. But both of these are wrong. An ordinary man who saw, for example, a rainbow would not, if persuaded that a rainbow is not a material thing, at once conclude that his senses were deceiving him; nor, when for instance he knows that the ship at sea on a clear day is much farther away than it looks, does he

conclude that he is not seeing a material thing (still less that he *is* seeing an immaterial ship). That is to say, there is no more a simple contrast between what the ordinary man believes when all is well (that he is 'perceiving material things') and when something is amiss (that his 'senses are deceiving him' and he is *not* 'perceiving material things') than there is between what he believes that he perceives ('material things') and what philosophers for their part are prepared to admit, whatever that may be. The ground is already being prepared for *two* bogus dichotomies.

3. Next, is it not rather delicately hinted in this passage that the plain man is really a bit naïve?[7] It 'does not normally occur' to him that his belief in 'the existence of material things' needs justifying—but perhaps it *ought* to occur to him. He has 'no doubt whatsoever' that he really perceives chairs and tables—but perhaps he ought to have a doubt or two and not be so easily 'satisfied.' That people are sometimes deceived by their senses 'does not lead him to suspect' that all may not be well—but perhaps a more reflective person *would* be led to suspect. Though ostensibly the plain man's position is here just being described, a little quiet undermining is already being effected by these turns of phrase.
4. But, perhaps more importantly, it is also implied, even taken for granted, that there is *room* for doubt and suspicion, whether or not the plain man feels any. The quotation from Locke, with which most people are said to agree, in fact contains a strong *suggestio falsi.* It suggests that when, for instance, I look at a chair a few yards in front of me in broad daylight, my view is that I have (*only*) as much certainty as I need and can get that there is a chair and that I see it. But in fact the plain man would regard doubt in such a case, not as far-fetched or over-refined or somehow unpractical, but as plain *nonsense*; he would say, quite correctly, 'Well, if that's not seeing a real chair then *I don't know what is.*' Moreover, though the plain man's alleged belief that his 'sense-perceptions' can 'in general' or 'now' be trusted is implicitly contrasted with the philosophers' view, it turns out that the philosophers' view is not just that his sense-perceptions *can't* be trusted 'now,' or 'in general,' or as often as he thinks; for apparently philosophers 'for the most part' really maintain that what the plain man believes to be the case is really *never* the case—'what, in their opinion, we directly perceive is *always* an object of a different kind.' The philosopher is not really going to argue that things go wrong more often than the unwary plain man supposes, but that in some sense or some way he is wrong all the time. So it is misleading to hint, not only that there is always room for doubt, but that the philosophers' dissent from the plain man is just a matter of degree; it is really not *that* kind of disagreement at all.
5. Consider next what is said here about deception. We recognize, it is said, that 'people are sometimes deceived by their senses,' though we think that, in general, our 'sense-perceptions' can 'be trusted.'

Now first, though the phrase 'deceived by our senses' is a common metaphor, it *is* a metaphor; and this is worth noting, for in what follows the same metaphor is frequently taken up by the expression 'veridical' and taken very seriously. In fact, of course, our senses are dumb—though Descartes and others speak of 'the testimony of the senses,' our senses do not *tell* us anything, true or false. The case is made much worse here by the unexplained introduction of a quite new creation, our 'sense-perceptions.' These entities, which of course don't really figure at all in the plain man's language or among his beliefs, are brought in with the implication that whenever we 'perceive' there is an *intermediate* entity *always* present and *informing* us about something *else*—the question is, can we or can't we trust what it says? Is it 'veridical'? But of course to state the case in this way is simply to soften up the plain man's alleged views for the subsequent treatment; it is preparing the way for, by practically attributing to *him,* the so-called philosophers' view.

Next, it is important to remember that talk of deception only *makes sense* against a background of general non-deception. (You can't fool all of the people all of the time.) It must be possible to *recognize* a case of deception by checking the odd case against more normal ones. If I say, 'Our petrol-gauge sometimes deceives us,' I am understood: though usually what it indicates squares with what we have in the tank, sometimes it doesn't—it sometimes points to two gallons when the tank turns out to be nearly empty. But suppose I say, 'Our crystal ball sometimes deceives us': this is puzzling,

because really we haven't the least idea what the 'normal' case—*not* being deceived by our crystal ball—would actually be.

The cases, again, in which a plain man might say he was 'deceived by his senses' are not at all common. In particular, he would *not* say this when confronted with ordinary cases of perspective, with ordinary mirror-images, or with dreams; in fact, when he dreams, looks down the long straight road, or at his face in the mirror, he is not, or at least is hardly ever, *deceived* at all. This is worth remembering in view of another strong *suggestio falsi*—namely, that when the philosopher cites as cases of 'illusion' all these and many other very common phenomena, he is either simply mentioning cases which the plain man already concedes as cases of 'deception by the senses,' or at any rate is only extending a bit what he would readily concede. In fact this is very far indeed from being the case.

And even so—even though the plain man certainly does not accept anything like so *many* cases as cases of being 'deceived by his senses' as philosophers seem to—it would certainly be quite wrong to suggest that he regards all the cases he *does* accept as being of just the same kind. The battle is, in fact, half lost already if this suggestion is tolerated. Sometimes the plain man would prefer to say that his senses were deceived rather than that he was deceived by his senses—the quickness of the hand deceives the eye, etc. But there is actually a great multiplicity of cases here, at least at the edges of which it is no doubt uncertain (and it would be typically scholastic to try to decide) just which are and which are not cases where the metaphor of being 'deceived by the senses' would naturally be employed. But surely even the plainest of men would want to distinguish (*a*) cases where the *sense-organ* is deranged or abnormal or in some way or other not functioning properly; (*b*) cases where the *medium*—or more generally, the conditions—of perception are in some way abnormal or off-colour; and (*c*) cases where a wrong inference is made or a wrong construction is put on things, e.g. on some sound that he hears. (Of course these cases do not exclude each other.) And then again there are the quite common cases of misreadings, mishearings, Freudian over-sights, etc., which don't seem to belong properly under any of these headings. That is to say, once again there is no neat and simple dichotomy between things going right and things going wrong; things may go wrong, as we really all know quite well, in lots of *different* ways—which don't have to be, and must not be assumed to be, classifiable in any general fashion.

Finally, to repeat here a point we've already mentioned, of course the plain man does *not* suppose that all the cases in which he is 'deceived by his senses' are alike in the particular respect that, in those cases, he is not 'perceiving material things,' or *is* perceiving something not real or not material. Looking at the Müller-Lyer diagram (in which, of two lines of equal length, one looks longer than the other), or at a distant village on a very clear day across a valley, is a very different kettle of fish from seeing a ghost or from having D.T.s and seeing pink rats. And when the plain man sees on the stage the Headless Woman, what he sees (and this *is* what he sees, whether he knows it or not) is not something 'unreal' or 'immaterial,' but a woman against a dark background with her head in a black bag. If the trick is well done, he doesn't (because it's deliberately made very difficult for him) properly size up what he sees, or see *what* it is; but to say this is far from concluding that he sees something *else*.

In conclusion, then, there is less than no reason to swallow the suggestions *either* that what the plain man believes that he perceives most of the time constitutes a *kind* of things (*sc.* 'material objects'), *or* that he can be said to recognize any other single *kind* of cases in which he is 'deceived.'[8] Now let us consider what it is that is said about philosophers.

Philosophers, it is said, 'are not, for the most part, prepared to admit that such objects as pens or cigarettes are ever directly perceived.' Now of course what brings us up short here is the word 'directly'—a great favourite among philosophers, but actually one of the less conspicuous snakes in the linguistic grass. We have here, in fact, a typical case of a word, which already has a very special use, being gradually stretched, without caution or definition or any limit, until it becomes, first perhaps obscurely metaphorical, but ultimately meaningless. One can't abuse ordinary language without paying for it.[9]

1. First of all, it is essential to realize that here the notion of perceiving *in*directly wears the trousers—'directly' takes whatever sense it has from the contrast with its opposite:[10] while 'indirectly' itself (*a*) has a use only in special cases, and also (*b*) has *different* uses in different cases—though that doesn't mean, of course, that there is not a good reason why we should

use the same word. We might, for example, contrast the man who saw the procession directly with the man who saw it *through a periscope;* or we might contrast the place from which you can watch the door directly with the place from which you can see it only *in the mirror. Perhaps* we might contrast seeing you directly with seeing, say, your shadow on the blind; and *perhaps* we might contrast hearing the music directly with hearing it relayed outside the concert-hall. However, these last two cases suggest two further points.

2. The first of these points is that the notion of not perceiving 'directly' seems most at home where, as with the periscope and the mirror, it retains its link with the notion of a kink in *direction.* It seems that we must not be looking *straight at* the object in question. For this reason seeing your shadow on the blind is a doubtful case; and seeing you, for instance, through binoculars or spectacles is certainly not a case of seeing you *indirectly* at all. For such cases as these last we have quite distinct contrasts and different expressions—'with the naked eye' as opposed to 'with a telescope,' 'with unaided vision' as opposed to 'with glasses on.' (These expressions, in fact, are much more firmly established in ordinary use than 'directly' is.)
3. And the other point is that, partly no doubt for the above reason, the notion of indirect perception is not naturally at home with senses other than sight. With the other senses there is nothing quite analogous with the 'line of vision.' The most natural sense of 'hearing indirectly,' of course, is that of being *told* something by an intermediary—a quite different matter. But do I hear a shout indirectly, when I hear the echo? If I touch you with a bargepole, do I touch you indirectly? Or if you offer me a pig in a poke, might I feel the pig indirectly—*through* the poke? And what smelling indirectly might be I have simply no idea. For this reason alone there seems to be something badly wrong with the question, 'Do we perceive things directly or not?,' where perceiving is evidently intended to cover the employment of *any* of the senses.
4. But it is, of course, for other reasons too extremely doubtful how far the notion of perceiving indirectly could or should be extended. Does it, or should it, cover the telephone, for instance? Or television? Or radar? Have we moved too far in these cases from the original metaphor? They at any rate satisfy what seems to be a necessary condition—namely, concurrent existence and concomitant variation as between what is perceived in the straightforward way (the sounds in the receiver, the picture and the blips on the screen) and the candidate for what we might be prepared to describe as being perceived indirectly. And this condition fairly clearly rules out as cases of indirect perception seeing photographs (which statically record scenes from the past) and seeing films (which, though not static, are not seen contemporaneously with the events thus recorded). Certainly, there *is* a line to be drawn somewhere. It is certain, for instance, that we should not be prepared to speak of indirect perception in *every* case in which we see something from which the existence (or occurrence) of something else can be inferred; we should *not* say we see the guns indirectly, if we see in the distance only the flashes of guns.
5. Rather differently, if we are to be seriously inclined to speak of something as being perceived indirectly, it seems that it has to be the kind of thing which we (sometimes at least) just perceive, or could perceive, or which—like the backs of our own heads—others could perceive. For otherwise we don't want to say that we perceive the thing *at all,* even indirectly. No doubt there are complications here (raised, perhaps, by the electron microscope, for example, about which I know little or nothing). But it seems clear that, in general, we should want to distinguish between seeing indirectly, e.g. in a mirror, what we might have just *seen,* and seeing signs (or effects), e.g. in a Wilson cloud-chamber, of something not itself perceptible at all. It would at least not come naturally to speak of the latter as a case of perceiving something indirectly.
6. And one final point. For reasons not very obscure, we always prefer in practice what might be called the *cash-value* expression to the 'indirect' metaphor. If I were to report that I see enemy ships indirectly, I should merely provoke the question what exactly I mean. 'I mean that I can see these blips on the radar screen'—'Well, why didn't you say so then?' (Compare 'I can see an unreal duck.'—'What on earth do

you mean?' 'It's a decoy duck'—'Ah, I see. Why didn't you say so at once?') That is, there is seldom if ever any particular point in actually saying 'indirectly' (or 'unreal'); the expression can cover too many rather different cases to be *just* what is wanted in any particular case.

Thus, it is quite plain that the philosophers' use of 'directly perceive,' whatever it may be, is not the ordinary, or any familiar, use; for in *that* use it is not only false but simply absurd to say that such objects as pens or cigarettes are never perceived directly. But we are given no explanation or definition of this new use[11]—on the contrary, it is glibly trotted out as if we were all quite familiar with it already. It is clear, too, that the philosophers' use, whatever it may be, offends against several of the canons just mentioned above—no restrictions whatever seem to be envisaged to any special circumstances or to any of the senses in particular, and moreover it seems that what we are to be said to perceive indirectly is *never*—is not the kind of thing which ever *could* be—perceived directly.

All this lends poignancy to the question Ayer himself asks, a few lines below the passage we have been considering: 'Why may we not say that we are directly aware of material things?' The answer, he says, is provided 'by what is known as the argument from illusion'; and this is what we must next consider. Just possibly the answer may help us to understand the question.

III

The primary purpose of the argument from illusion is to induce people to accept 'sense-data' as the proper and correct answer to the question what they perceive on certain *abnormal, exceptional* occasions; but in fact it is usually followed up with another bit of argument intended to establish that they *always* perceive sense-data. Well, what is the argument?

In Ayer's statement[12] it runs as follows. It is 'based on the fact that material things may present different appearances to different observers, or to the same observer in different conditions, and that the character of these appearances is to some extent causally determined by the state of the conditions and the observer.' As illustrations of this alleged fact Ayer proceeds to cite perspective ('a coin which looks circular from one point of view may look elliptical from another'); refraction ('a stick which normally appears straight looks bent when it is seen in water'); changes in colour-vision produced by drugs ('such as mescal'); mirror-images; double vision; hallucination; apparent variations in tastes; variations in felt warmth ('according as the hand that is feeling it is itself hot or cold'); variations in felt bulk ('a coin seems larger when it is placed on the tongue than when it is held in the palm of the hand'); and the oft-cited fact that 'people who have had limbs amputated may still continue to feel pain in them.'

He then selects three of these instances for detailed treatment. First, refraction—the stick which normally 'appears straight' but 'looks bent' when seen in water. He makes the 'assumptions' (*a*) that the stick does not *really change its shape* when it is placed in water, and (*b*) that it *cannot be* both crooked and straight.[13] He then concludes ('it follows') that 'at least one of the *visual appearances* of the stick is *delusive*.' Nevertheless, even when 'what we see is not the *real quality* of a *material thing*, it is supposed that we are still seeing something'—and this something is to be called a 'sense-datum.' A sense-datum is to be 'the object of which we are *directly* aware, in perception, if it is not *part* of any *material thing*.' (The italics are mine throughout this and the next two paragraphs.)

Next, mirages. A man who sees a mirage, he says, is 'not perceiving any material thing; for the oasis which he thinks he is perceiving *does not exist*.' But 'his *experience* is not an experience of nothing'; thus 'it is said that he is experiencing sense-data, which are similar in character to what he would be experiencing if he were seeing a real oasis, but are delusive in the sense that *the material thing which they appear to present* is not *really there*.'

Lastly, reflections. When I look at myself in a mirror 'my body *appears to be* some distance behind the glass'; but it cannot actually be in two places at once; thus, my perceptions in this case 'cannot all be *veridical*.' But I do see *something*, and if 'there really is no such material thing as my body in the place where it appears to be, what is it that I am seeing?' Answer—a sense-datum. Ayer adds that 'the same conclusion may be reached by taking any other of my examples.'

Now I want to call attention, first of all, to the name of this argument—the 'argument from *illusion*,' and to the fact that it is produced as establishing the conclusion that some at least of our 'perceptions' are *delusive*. For in this there are two clear implications—(*a*) that all the cases

cited in the argument are cases of *illusions*; and (*b*) that *illusion* and *delusion* are the same thing. But both of these implications, of course, are quite wrong; and it is by no means unimportant to point this out, for, as we shall see, the argument trades on confusion at just this point.

What, then, would be some genuine examples of illusion? (The fact is that hardly any of the cases cited by Ayer is, at any rate without stretching things, a case of illusion at all.) Well, first, there are some quite clear cases of *optical* illusion—for instance the case we mentioned earlier in which, of two lines of equal length, one is made to look longer than the other. Then again there are illusions produced by professional 'illusionists,' conjurors—for instance the Headless Woman on the stage, who is made to look headless, or the ventriloquist's dummy which is made to appear to be talking. Rather different—not (usually) produced on purpose—is the case where wheels rotating rapidly enough in one direction may look as if they were rotating quite slowly in the opposite direction. Delusions, on the other hand, are something altogether different from this. Typical cases would be delusions of persecution, delusions of grandeur. These are primarily a matter of grossly disordered beliefs (and so, probably, behaviour) and may well have nothing in particular to do with perception.[14] But I think we might also say that the patient who sees pink rats has (suffers from) delusions— particularly, no doubt, if, as would probably be the case, he is not clearly aware that his pink rats aren't real rats.[15]

The most important differences here are that the term 'an illusion' (in a perceptual context) does not suggest that something totally unreal is *conjured up*—on the contrary, there just is the arrangement of lines and arrows on the page, the woman on the stage with her head in a black bag, the rotating wheels; whereas the term 'delusion' *does* suggest something totally unreal, not really there at all. (The convictions of the man who has delusions of persecution can be *completely* without foundation.) For this reason delusions are a much more serious matter—something is really wrong, and what's more, wrong *with* the person who has them. But when I see an optical illusion, however well it comes off, there is nothing wrong with me personally, the illusion is not a little (or a large) peculiarity or idiosyncrasy of my own; it is quite public, anyone can see it, and in many cases standard procedures can be laid down for producing it. Furthermore, if we are not actually to be taken in, we need to be *on our guard*; but it is no use to tell the sufferer from delusions to be on his guard. He needs to be cured.

Why is it that we tend—if we do—to confuse illusions with delusions? Well, partly, no doubt the terms are often used loosely. But there is also the point that people may have, without making this explicit, different views or theories about the facts of some cases. Take the case of seeing a ghost, for example. It is not generally known, or agreed, what seeing ghosts *is*. Some people think of seeing ghosts as a case of something being conjured up, perhaps by the disordered nervous system of the victim; so in their view seeing ghosts is a case of delusion. But other people have the idea that what is called seeing ghosts is a case of being taken in by shadows, perhaps, or reflections, or a trick of the light—that is, they assimilate the case in their minds to illusion. In this way, seeing ghosts, for example, may come to be labelled sometimes as 'delusion,' sometimes as 'illusion'; and it may not be noticed that it makes a difference which label we use. Rather, similarly, there seem to be different doctrines in the field as to what mirages are. Some seem to take a mirage to be a vision conjured up by the crazed brain of the thirsty and exhausted traveler (delusion), while in other accounts it is a case of atmospheric refraction, whereby something below the horizon is made to appear above it (illusion). (Ayer, you may remember, takes the delusion view, although he cites it along with the rest as a case of illusion. He says not that the oasis appears to be where it is not, but roundly that 'it does not exist.')

The way in which the 'argument from illusion' positively trades on not distinguishing illusions from delusions is, I think, this. So long as it is being suggested that the cases paraded for our attention are cases of *illusion*, there is the implication (from the ordinary use of the word) that there really is something there that we perceive. But then, when these cases begin to be quietly called delusive, there comes in the very different suggestion of something being conjured up, something unreal or at any rate 'immaterial.' These two implications taken together may then subtly insinuate that in the cases cited there really is something that we are perceiving, but that this is an immaterial something; and this insinuation, even if not conclusive by itself, is certainly well calculated to edge us a little closer towards just the position where the sense-datum theorist wants to have us.

So much, then—though certainly there could be a good deal more—about the differences between illusions and delusions and the reasons for not obscuring them. Now let us look briefly at some of the other cases Ayer lists. Reflections, for instance. No doubt you *can* produce illusions with mirrors, suitably disposed. But is just *any* case of seeing something in a mirror an illusion, as he implies? Quite obviously not. For seeing things in mirrors is a perfectly *normal* occurrence, completely familiar, and there is usually no question of anyone being taken in. No doubt, if you're an infant or an aborigine and have never come across a mirror before, you may be pretty baffled, and even visibly perturbed, when you do. But is that a reason why the rest of us should speak of illusion here? And just the same goes for the phenomena of perspective—again, one *can* play tricks with perspective, but in the ordinary case there is no question of illusion. That a round coin should 'look elliptical' (in one sense) from some points of view is exactly what we expect and what we normally find; indeed, we should be badly put out if we ever found this not to be so. Refraction again—the stick that looks bent in water—is far too familiar a case to be properly called a case of illusion. We may perhaps be prepared to agree that the stick looks bent; but then we can see that it's partly submerged in water, so that is exactly how we should expect it to look.

It is important to realize here how familiarity, so to speak, takes the edge off illusion. Is the cinema a case of illusion? Well, just possibly the first man who ever saw moving pictures may have felt inclined to say that here was a case of illusion. But in fact it's pretty unlikely that even he, even momentarily, was actually taken in; and by now the whole thing is so ordinary a part of our lives that it never occurs to us even to raise the question. One might as well ask whether producing a photograph is producing an illusion—which would plainly be just silly.

Then we must not overlook, in all this talk about illusions and delusions, that there are plenty of more or less unusual cases, not yet mentioned, which certainly aren't either. Suppose that a proof-reader makes a mistake—he fails to notice that what ought to be 'causal' is printed as 'casual'; does he have a delusion? Or is there an illusion before him? Neither, of course; he simply *misreads*. Seeing after-images, too, though not a particularly frequent occurrence and not just an ordinary case of seeing, is neither seeing illusions nor having delusions. And what about dreams? Does the dreamer see illusions? Does he have delusions? Neither; dreams are *dreams*.

Let us turn for a moment to what Price has to say about illusions. He produces,[16] by way of saying 'what the term 'illusion' means,' the following 'provisional definition': 'An illusory sense-datum of sight or touch is a sense-datum which is such that we tend to take it to be part of the surface of a material object, but if we take it so we are wrong.' It is by no means clear, of course, what this dictum itself means; but still, it seems fairly clear that the definition doesn't actually fit all the cases of illusion. Consider the two lines again. Is there anything here which we tend to take, wrongly, to be part of the surface of a material object? It doesn't seem so. We just see the two lines, we don't think or even tend to think that we see anything else, we aren't even raising the question whether anything is or isn't 'part of the surface' of— what, anyway? the lines? the page?—the trouble is just that one line looks longer than the other, though it isn't. Nor surely, in the case of the Headless Woman, is it a question whether anything is or isn't part of her surface; the trouble is just that she looks as if she had no head.

It is noteworthy, of course, that, before he even begins to consider the 'argument from illusion,' Price has already incorporated in this 'definition' the idea that in such cases there is something to be seen *in addition to* the ordinary things—which is part of what the argument is commonly used, and not uncommonly taken, to *prove*. But this idea surely has no place in an attempt to say what 'illusion' *means*. It comes in again, improperly I think, in his account of perspective (which incidentally he also cites as a species of illusion)—'a distant hillside which is full of protuberances, and slopes upwards at quite a gentle angle, will appear flat and vertical. . . . This means that the sense-datum, the colour-expanse which we sense, actually *is* flat and vertical.' But why should we accept this account of the matter? Why should we say that there is *anything* we see which *is* flat and vertical, though not 'part of the surface' of any material object? To speak thus is to assimilate all such cases to cases of delusion, where there *is* something not 'part of any material thing.' But we have already discussed the undesirability of this assimilation.

Next, let us have a look at the account Ayer himself gives of some at least of the cases he cites. (In fairness we must remember here that

Ayer has a number of quite substantial reservations of his own about the merits and efficacy of the argument from illusion, so that it is not easy to tell just how seriously he intends his exposition of it to be taken; but this is a point we shall come back to.)

First, then, the familiar case of the stick in water. Of this case Ayer says (*a*) that since the stick looks bent but is straight, 'at least one of the visual appearances of the stick is *delusive*'; and (*b*) that 'what we see [directly anyway] is not the real quality of [a few lines later, not part of] a material thing.' Well now: does the stick 'look bent' to begin with? I think we can agree that it does, we have no better way of describing it. But of course it does *not* look *exactly* like a bent stick, a bent stick out of water—at most, it may be said to look rather like a bent stick partly immersed *in* water. After all, we can't help seeing the water the stick is partly immersed in. So exactly what in this case is supposed to be *delusive*? What is wrong, what is even faintly surprising, in the idea of a stick's being straight but looking bent sometimes? Does anyone suppose that if something is straight, then it jolly well has to *look* straight at all times and in all circumstances? Obviously no one seriously supposes this. So what mess are we supposed to get into here, what is the difficulty? For of course it has to be suggested that there *is* a difficulty—a difficulty, furthermore, which calls for a pretty radical solution, the introduction of sense-data. But what is the problem we are invited to solve in this way?

Well, we are told, in this case you are seeing *something*; and what is this something 'if it is not part of any material thing'? But this question is, really, completely mad. The straight part of the stick, the bit not under water, is presumably part of a material thing; don't we see that? And what about the bit *under* water?—we can see that too. We can see, come to that, the water itself. In fact what we see is *a stick partly immersed in water*; and it is particularly extraordinary that this should appear to be called in question—that a question should be raised about *what* we are seeing—since this, after all, is simply the description of the situation with which we started. It was, that is to say, agreed at the start that we were looking at a stick, a 'material thing,' part of which was under water. If, to take a rather different case, a church were cunningly camouflaged so that it looked like a barn, how could any serious question be raised about what we see when we look at it? We see, of course, *a church* that now *looks like a barn.* We do *not* see an immaterial barn, an immaterial church, or an immaterial anything else. And what in this case could seriously tempt us to say that we do?

Notice, incidentally, that in Ayer's description of the stick-in-water case, which is supposed to be prior to the drawing of any philosophical conclusions, there has already crept in the unheralded but important expression 'visual appearances'—it is, of course, ultimately to be suggested that all we *ever* get when we see is a visual appearance (whatever that may be).

Consider next the case of my reflection in a mirror. My body, Ayer says, 'appears to be some distance behind the glass'; but as it's in front, it can't really be behind the glass. So what am I seeing? A sense-datum. What about this? Well, once again, although there is no objection to saying that my body 'appears to be some distance behind the glass,' in saying this we must remember what sort of situation we are dealing with. It does not 'appear to be' there in a way which might tempt me (though it might tempt a baby or a savage) to go round the back and look for it, and be astonished when this enterprise proved a failure. (To say that A is *in* B doesn't always mean that if you open B you will find A, just as to say that A is *on* B doesn't always mean that you could pick it off—consider 'I saw my face in the mirror,' 'There's a pain in my toe,' 'I heard him on the radio,' 'I saw the image on the screen,' etc. Seeing something in a mirror is not like seeing a bun in a shop-window.) But does it follow that, since my body is not actually located behind the mirror, I am not seeing a material thing? Plainly not. For one thing, I can see the mirror (nearly always anyway). I can see my own body 'indirectly,' sc. in the mirror. I can also see the reflection of my own body or, as some would say, a mirror-image. And a mirror-image (if we choose this answer) is not a 'sense-datum'; it can be photographed, seen by any number of people, and so on. (Of course there is no question here of either illusion or delusion.) And if the question is pressed, what actually *is* some distance, five feet say, behind the mirror, the answer is, not a sense-datum, but some region of the adjoining room.

The mirage case—at least if we take the view, as Ayer does, that the oasis the traveler thinks he can see 'does not exist'—is significantly more amenable to the treatment it is given. For here we are supposing the man to be genuinely deluded, he is *not* 'seeing a material thing.'[17] We don't actually have to say, however, even

here that he is 'experiencing sense-data'; for though, as Ayer says above, 'it is convenient to give a name' to what he is experiencing, the fact is that it already has a name—a *mirage*. Again, we should be wise not to accept too readily the statement that what he is experiencing is '*similar in character* to what he would be experiencing if he were seeing a real oasis.' For is it at all likely, really, to be very similar? And, looking ahead, if we were to concede this point we should find the concession being used against us at a later stage—namely, at the stage where we shall be invited to agree that we see sense-data always, in normal cases too.

NOTES

1. Macmillan, 1940.
2. Methuen, 1932.
3. Penguin Books, 1953.
4. The case of 'universal' and 'particular,' or 'individual,' is similar in some respects though of course not in all. In philosophy it is often good policy, where one member of a putative pair falls under suspicion, to view the more innocent-seeming party suspiciously as well.
5. Ayer, op. cit., pp. 1–2.
6. Compare Price's list on p. I of *Perception*—'chairs and tables, cats and rocks'—though he complicates matters by adding 'water' and 'the earth.' See also p. 280, on 'physical objects,' 'visuo-tactual solids.'
7. Price, op. cit., p. 26, says that he *is* naive, though it is not, it seems, certain that he is actually a Naive Realist.
8. I am not denying that cases in which things go wrong *could* be lumped together under some single name. A single name might in itself be innocent enough, provided its use was not taken to imply either (*a*) that the cases were all alike, or (*b*) that they were all in certain ways alike. What matters is that the facts should not be pre-judged and (therefore) neglected.
9. Especially if one abuses it without realizing what one is doing. Consider the trouble caused by unwitting stretching of the word 'sign,' so as to yield—apparently—the conclusion that, when the cheese is in front of our noses, we see signs of cheese.
10. Compare, in this respect, 'real,' 'proper,' 'free,' and plenty of others. 'It's real'—what exactly are you saying it isn't? 'I wish we had a proper stair-carpet'—what are you complaining of in the one you've got? (That it's *im*proper?) 'Is he free?'—well, what have you in mind that he might be instead? In prison? Tied up in prison? Committed to a prior engagement?
11. Ayer takes note of this, rather belatedly, on pp. 60–61.
12. Ayer, op. cit., pp. 3–5.
13. It is not only strange, but also important, that Ayer calls these 'assumptions.' Later on he is going to take seriously the notion of denying at least one of them, which he could hardly do if he had recognized them here as the plain and incontestable facts that they are.
14. The latter point holds, of course, for *some* uses of 'illusion' too; there are the illusions which some people (are said to) lose as they grow older and wiser.
15. Cp. the white rabbit in the play called *Harvey*.
16. *Perception,* p. 27.
17. Not even 'indirectly,' no such thing is 'presented.' Doesn't this seem to make the case, though more amenable, a good deal less useful to the philosopher? It's hard to see how normal cases could be said to be *very like* this.

54 The 'Sensation' as a Unit of Experience

Maurice Merleau-Ponty

At the outset of the study of perception, we find in language the notion of sensation, which seems immediate and obvious: I have a sensation of redness, of blueness, of hot or cold. It will, however, be seen that nothing could in fact be more confused, and that because they accepted it readily, traditional analyses missed the phenomenon of perception.

I might in the first place understand by sensation the way in which I am affected and the experiencing of a state of myself. The greyness which, when I close my eyes, surrounds me, leaving no distance between me and it, the sounds that encroach on my drowsiness and hum 'in my head' perhaps give some indication of what pure sensation might be. I might be said

Excerpted from *The Phenomenenology of Perception* (Humanities Press, 1962).

to have sense-experience (*sentir*) precisely to the extent that I coincide with the sensed, that the latter ceases to have any place in the objective world, and that it signifies nothing for me. This entails recognizing that sensation should be sought on the hither side of any qualified content, since red and blue, in order to be distinguishable as two colours, must already form some picture before me, even though no precise place be assigned to them, and thus cease to be part of myself. Pute sensation will be the experience of an undifferentiated, instantaneous, dotlike impact. It is unnecessary to show, since authors are agreed on it, that this notion corresponds to nothing in our experience, and that the most rudimentary *factual perceptions* that we are acquainted with, in creatures such as the ape or the hen, have a bearing on relationships and not on any absolute terms.[1] But this does not dispose of the question as to why we feel justified *in theory* in distinguishing within experience a layer of 'impression.' Let us imagine a white patch on a homogeneous background. All the points in the patch have a certain 'function' in common, that of forming themselves into a 'shape.' The colour of the shape is more intense, and as it were more resistant than that of the background; the edges of the white patch 'belong' to it, and are not part of the background although they adjoin it: the patch appears to be placed on the background and does not break it up. Each part arouses the expectation of more than it contains, and this elementary perception is therefore already charged with a *meaning*. But if the shape and the background, as a whole, are not sensed, they must be sensed, one may object, in each of their points. To say this is to forget that each point in its turn can be perceived only as a figure on a background. When Gestalt theory informs us that a figure on a background is the simplest sense-given available to us, we reply that this is not a contingent characteristic of factual perception, which leaves us free, in an ideal analysis, to bring in the notion of impressions. It is the very definition of the phenomenon of perception, that without which a phenomenon cannot be said to be perception at all. The perceptual 'something' is always in the middle of something else, it always forms part of a 'field.' A really homogeneous area offering *nothing to be* cannot be given to *any perception*. The structure of actual perception alone can teach us what perception is. The pure impression is, therefore, not only undiscoverable, but also imperceptible and so inconceivable as an instant of perception. If it is introduced, it is because instead of attending to the experience of perception, we overlook it in favour of the object perceived. A visual field is not made up of limited views. But an object seen is made up of bits of matter, and spatial points are external to each other. An isolated datum of perception is inconceivable, at least if we do the mental experiment of attempting to perceive such a thing. But in the world there are either isolated objects or a physical void.

I shall therefore give up any attempt to define sensation as pure impression. Rather, to see is to have colours or lights, to hear is to have sounds, to sense (*sentir*) is to have qualities. To know what sense-experience is, then, is it not enough to have seen a red or to have heard an A? But red and green are not sensations, they are the sensed (*sensibles*), and quality is not an element of consciousness, but a property of the object. Instead of providing a simple means of delimiting sensations, if we consider it in the experience itself which evinces it, the quality is as rich and mysterious as the object, or indeed the whole spectacle, perceived. This red patch which I see on the carpet is red only in virtue of a shadow which lies across it, its quality is apparent only in relation to the play of light upon it, and hence as an element in a spatial configuration. Moreover the colour can be said to be there only if it occupies an area of a certain size, too small an area not being describable in these terms. Finally this red would literally not be the same if it were not the 'woolly red' of a carpet.[2] Analysis, then, discovers in each quality meanings which reside in it. It may be objected that this is true only of the qualities which form part of our actual experience, which are overlaid with a body of knowledge, and that we are still justified in conceiving a 'pure quality' which would set limits to a pure sensation. But as we have just seen, this pure sensation would amount to no sensation, and thus to not feeling at all. The alleged self-evidence of sensation is not based on any testimony of consciousness, but on widely held prejudice. We think we know perfectly well what 'seeing,' 'hearing,' 'sensing' are, because perception has long provided us with objects which are coloured or which emit sounds. When we try to analyse it, we transpose these objects into consciousness. We commit what psychologists call 'the experience error,' which means that what we know to be in things themselves we immediately take as being in our consciousness of them. We make perception out of things perceived. And since perceived things themselves are obviously

accessible only through perception, we end by understanding neither. We are caught up in the world and we do not succeed in extricating ourselves from it in order to achieve consciousness of the world. If we did we should see that the quality is never experienced immediately, and that all consciousness is consciousness of something. Nor is this 'something' necessarily an identifiable object. There are two ways of being mistaken about quality: one is to make it into an element of consciousness, when in fact it is an object *for* consciousness, to treat it as an incommunicable impression, whereas it always has a meaning; the other is to think that this meaning and this object, at the level of quality, are fully developed and determinate. The second error, like the first, springs from our prejudice about the world. Suppose we construct, by the use of optics and geometry, that bit of the world which can at any moment throw its image on our retina. Everything outside its perimeter, since it does not reflect upon any sensitive area, no more affects our vision than does light falling on our closed eyes. We ought, then, to perceive a segment of the world precisely delimited, surrounded by a zone of blackness, packed full of qualities with no interval between them, held together by definite relationships of size similar to those lying on the retina. The fact is that experience offers nothing like this, and we shall never, using the world as our starting-point, understand what a *field of vision* is. Even if it is possible to trace out a perimeter of vision by gradually approaching the centre of the lateral stimuli, the results of such measurement vary from one moment to another, and one never manages to determine the instant when a stimulus once seen is seen no longer. The region surrounding the visual field is not easy to describe, but what is certain is that it is neither black nor grey. There occurs here an *indeterminate vision*, a *vision of something or other*, and, to take the extreme case, what is behind my back is not without some element of visual presence. The two straight lines in Müller-Lyer's optical illusion (Fig. 1) are neither of equal nor unequal length; it is only in the objective world that this question arises.[3] The visual field is that strange zone in which contradictory notions jostle each other because the objects—the straight lines of Müller-Lyer—are not, in that field, assigned to the realm of being, in which a comparison would be possible, but each is taken in its private context as if it did not belong to the same universe as the other. Psychologists have for a long time taken great care to overlook these phenomena. In the world taken in itself everything is determined. There are many unclear sights, as for example a landscape on a misty day, but then we always say that no real landscape is in itself unclear. It is so only for us. The object, psychologists would assert, is never ambiguous, but becomes so only through our inattention. The bounds of the visual field are not themselves variable, and there is a moment when the approaching object begins absolutely to be seen, but we do not 'notice' it. But the notion of attention, as we shall show more fully, is supported by no evidence provided by consciousness. It is no more than an auxiliary hypothesis, evolved to save the prejudice in favour of an objective world. We must recognize the indeterminate as a positive phenomenon. It is in this atmosphere that quality arises. Its meaning is an equivocal meaning; we are concerned with an expressive value rather than with logical signification. The determinate quality by which empiricism tried to define sensation is an object, not an element, of consciousness, indeed it is the very lately developed object of scientific consciousness. For these two reasons, it conceals rather than reveals subjectivity.

Figure 1

The two definitions of sensation which we have just tried out were only apparently direct. We have seen that they were based on the object perceived. In this they were in agreement with common sense, which also identifies the sensible by the objective conditions which govern it. The visible is what is seized upon *with* the eyes, the sensible is what is seized on *by* the senses. Let us follow up the idea of sensation on this basis,[4] and see what becomes of this 'by' and this 'with,' and the notion of sense-organ, in the first-order thinking constituted by science. Having shown that there is no experience of sensation, do we at least find, in its causes and objective origins, any reasons for retaining it as an explanatory concept? Physiology, to which the psychologist turns as to a higher court of appeal, is in the same predicament as psychology. It too first situates its object in the world and treats it as a bit of extension. *Behaviour* is thus hidden by the reflex, the elaboration and patterning of stimuli, by a longitudinal theory

of nervous functioning, which establishes a theoretical correspondence between each element of the situation and an element of the reaction.[5] As in the case of the reflex are theory, physiology of perception begins by recognizing an anatomical path leading from a *receive* through a definite *transmitter* to a recording station,[6] equally specialized. The objective world being given, it is assumed that it passes on to the sense-organs messages which must be registered, then deciphered in such a way as to reproduce in us the original text. Hence we have in principle a point-by-point correspondence and constant connection between the stimulus and the elementary perception. But this 'constancy hypothesis'[7] conflicts with the data of consciousness, and the very psychologists who accept it recognize its purely theoretical character.[8] For example, the intensity of a sound under certain circumstances lowers its pitch; the addition of auxiliary lines makes two figures unequal which are objectively equal;[9] a coloured area appears to be the same colour over the whole of its surface, whereas the chromatic thresholds of the different parts of the retina ought to make it red in one place, orange somewhere else, and in certain cases colourless.[10] Should these cases in which the phenomenon does not correspond to the stimulus be retained within the framework of the law of constancy, and explained by additional factors—attention and judgement—or must the law itself be jettisoned? When red and green, presented together, give the result grey, it is conceded that the central combination of stimuli can immediately give rise to a different sensation from what the objective stimuli would lead us to expect. When the apparent size of an object varies with its apparent distance, or its apparent colour with our recollections of the object, it is recognized that 'the sensory processes are not immune to central influences.'[11] In this case, therefore, the 'sensible' cannot be defined as the immediate effect of an external stimulus. Cannot the same conclusion be drawn from the first three examples we have mentioned? If attention, more precise instructions, rest or prolonged practice finally bring perception into line with the law of constancy, this does not prove the law's universal validity, for, in the examples quoted, the first appearance possessed a sensory character just as incontestable as the final results obtained. So the question is whether attentive perception, the subject's concentration on one point of the visual field—for example, the 'analytic perception' of the two main lines in Müller-Lyer's optical illusion—do not, instead of revealing the 'normal sensation,' substitute a special set-up for the original phenomenon.[12] The law of constancy cannot avail itself, against the testimony of consciousness, of any crucial experiment in which it is not already implied, and wherever we believe that we are establishing it, it is already presupposed.[13] If we turn back to the phenomena, they show us that the apprehension of a quality, just as that of size, is bound up with a whole perceptual context, and that the stimuli no longer furnish us with the indirect means we were seeking of isolating a layer of immediate impressions. But when we look for an 'objective' definition of sensation, it is not only the physical stimulus which slips through our fingers. The sensory apparatus, as conceived by modern physiology, is no longer fitted for the rôle of 'transmitter' cast for it by traditional science. Non-cortical lesions of the apparatus of touch no doubt lessen the concentration of points sensitive to heat and cold, or pressure, and diminish the sensitivity of those that remain. But if, to the injured system, a sufficiently extensive stimulus be applied, the specific sensations reappear. The raising of the thresholds is compensated by a more vigorous movement of the hand.[14]

One can discern, at the rudimentary stage of sensibility, a working together on the part of partial stimuli and a collaboration of the sensory with the motor system which, in a variable physiological constellation, keeps sensation constant, and rules out any definition of the nervous process as the simple transmission of a given message. The destruction of sight, wherever the injuries be sustained, follows the same law: all colours are affected in the first place,[15] and lose their saturation. Then the spectrum is simplified, being reduced to four and soon to two colours; finally a grey monochrome stage is reached, although the pathological colour is never identifiable with any normal one. Thus in central as in peripheral lesions 'the loss of nervous substance results not only in a deficiency of certain qualities, but in the change to a less differentiated and more primitive structure.'[16] Conversely, normal functioning must be understood as a process of integration in which the text of the external world is not so much copied, as composed. And if we try to seize 'sensation' within the perspective of the bodily phenomena which pave the way to it, we find not a psychic individual, a function of certain known variables, but a formation already bound up with a larger whole, already endowed with a meaning, distinguishable only in degree

from the more complex perceptions, and which therefore gets us no further in our attempt to delimit pure sensation. There is no physiological definition of sensation, and more generally there is no physiological psychology which is autonomous, because the physiological event itself obeys biological and psychological laws. For a long time it was thought that peripheral conditioning was the surest method of identifying 'elementary' psychic functions, and of distinguishing them from 'superior' functions less strictly bound up with the bodily substructure. A closer analysis, however, reveals that the two kinds of function overlap. The elementary is no longer that which by addition will cumulatively constitute the whole, nor is it a mere occasion for the whole to constitute itself. The elementary event is already invested with meaning, and the higher function will bring into being only a more integrated mode of existence or a more valid adaptation, by using and sublimating the subordinate operations. Conversely, 'sense-experience is a vital process, no less than procreation, breathing or growth.'[17] Psychology and physiology are no longer, then, two parallel sciences, but two accounts of behaviour, the first concrete, the second abstract.[18] We said that when the psychologist asks the physiologist for a definition of sensation 'in causal terms,' he encounters once more on this new ground his familiar difficulties, and now we can see why. The physiologist for his part has to rid himself of the realistic prejudice which all the sciences borrow from common sense, and which hampers them in their development. The changed meaning of the terms 'elementary' and 'more advanced' in modern physiology proclaims a changed philosophy.[19] The scientist too must learn to criticize the idea of an external world in itself, since the facts themselves prompt him to abandon that of the body as a transmitter of messages. The sensible is what is apprehended *with* the senses, but now we know that this 'with' is not merely instrumental, that the sensory apparatus is not a conductor, that even on the periphery the physiological impression is involved in relations formerly considered central.

Once more, reflection—even the second-order reflection of science—obscures what we thought was clear. We believed we knew what feeling, seeing and hearing were, and now these words raise problems. We are invited to go back to the experiences to which they refer in order to redefine them. The traditional notion of sensation was not a concept born of reflection, but a late product of thought directed towards objects, the last element in the representation of the world, the furthest removed from its original source, and therefore the most unclear. Inevitably science, in its general effort towards objectification, evolved a picture of the human organism as a physical system undergoing stimuli which were themselves identified by their physicochemical properties, and tried to reconstitute actual perception[20] on this basis, and to close the circle of scientific knowledge by discovering the laws governing the production of knowledge itself, by establishing an objective science of subjectivity.[21] But it is also inevitable that this attempt should fail. If we return to the objective investigations themselves, we first of all discover that the conditions external to the sensory field do not govern it part for part, and that they exert an effect only to the extent of making possible a basic pattern—which is what Gestalt theory makes clear. Then we see that within the organism the structure depends on variables such as the biological meaning of the situation, which are no longer physical variables, with the result that the whole eludes the well-known instruments of physico-mathematical analysis, and opens the way to another type of intelligibility.[22] If we now turn back, as is done here, towards perceptual experience, we notice that science succeeds in constructing only a semblance of subjectivity: it introduces sensations which are things, just where experience shows that there are meaningful patterns; it forces the phenomenal universe into categories which make sense only in the universe of science. It requires that two perceived lines, like two real lines, should be equal or unequal, that a perceived crystal should have a definite number of sides,[23] without realizing that the perceived, by its nature, admits of the ambiguous, the shifting, and is shaped by its context. In Müller-Lyer's illusion, one of the lines ceases to be equal to the other without becoming 'unequal': it becomes 'different.' That is to say, an isolated, objective line, and the same line taken in a figure, cease to be, for perception, 'the same.' It is identifiable in these two functions only by analytic perception, which is not natural. In the same way the perceived contains gaps which are not mere 'failures to perceive.' I may, through sight or touch, recognize a crystal as having a 'regular' shape without having, even tacitly, counted its sides. I may be familiar with a face without ever having perceived the colour of the eyes in themselves. The theory of sensation, which builds up all knowledge out of determinate qualities,

offers us objects purged of all ambiguity, pure and absolute, the ideal rather than the real themes of knowledge: in short, it is compatible only with the lately developed superstructure of consciousness. That is where 'the idea of sensation is approximately realized.'[24]

The images which instinct projects before it, those which tradition recreates in each generation, or simply dreams, are in the first place presented on an equal footing with genuine perceptions, and gradually, by critical labour, the true, present and explicit perception is distinguished from phantasms. The word perception indicates a *direction* rather than a primitive function.[25] It is known that the uniformity of apparent size of objects at different distances, or of their colour in different lights, is more perfect in children than in adults.[26] It follows that perception is more strictly bound up with the local stimulus in its developed than in its undeveloped state, and more in conformity with the theory of sensation in the adult than in the child. It is like a net with its knots showing up more and more clearly.[27] 'Primitive thought' has been pictured in a way which can be understood only if the responses of primitive people, their pronouncements and the sociologists' interpretations are related to the fund of perceptual experience which they are all trying to translate.[28] It is sometimes the adherence of the perceived object to its context, and, as it were, its viscosity, sometimes the presence in it of a positive indeterminate which prevents the spatial, temporal and numerical wholes from becoming articulated into manageable, distinct and identifiable terms. And it is this pre-objective realm that we have to explore in ourselves if we wish to understand sense experience.

NOTES

1. See *La Structure du Comportement*, pp. 142 and ff.
2. J. P. Sartre. *L'Imaginaire*, p. 241.
3. Koffka, *Psychologie*, p. 530.
4. There is no justification for dodging the issue, as does Jaspers, for example (*Zur Analyse der Trugwahrnehmungen*) by setting up in opposition, on the one hand a descriptive psychology which 'understands' phenomena, and on the other an explanatory psychology, which concerns itself with their origin. The psychologist always sees consciousness as placed in the body in the midst of the world, and for him the series stimulus-impression-perception is a sequence of events at the end of which perception begins. Each consciousness is born in the world and each perception is a new birth of consciousness. In this perspective the 'immediate' data of perception can always be challenged as mere appearances and as complex products of an origin. The descriptive method can acquire a genuine claim only from the transcendental point of view. But, even from this point of view, the problem remains as to how consciousness perceives itself or appears to itself as inserted in a nature. For the philosopher, as for the psychologist, there is therefore always a problem of origins, and the only method possible is to follow, in its scientific development, the causal explanation in order to make its meaning quite clear, and assign to it its proper place in the body of truth. That is why there will be found no *refutation*, but only an effort to understand the difficulties peculiar to causal thinking.
5. See *La Structure du Comportement*, chap. I.
6. We are translating roughly the series 'Empfänger-Übermittler-Empfinder,' spoken of by J. Stein, *Über die Veränderung der Sinnesleistungen und die Entstehung von Trugwahrnehmungen*, p. 351.
7. Koehler, *Über unbemerkte Empfindungen und Urteilstäuschungen*.
8. Stumpf does so explicitly. Cf. Koehler, ibid., p. 54.
9. Koehler, ibid., pp. 57–8, cf. pp. 58–66.
10. R. Déjean, *Les Conditions objectives de la Perception visuelle*, pp. 60 and 83.
11. Stumpf, quoted by Koehler, ibid., p. 58.
12. Koehler, ibid., pp. 58–63.
13. It is only fair to add that this is true of all theories, and that nowhere is there a crucial experiment. For the same reason the constancy hypothesis cannot be completely refuted on the basis of induction. It is discredited because it overlooks phenomena and does not permit any understanding of them. To discern them and to pass judgement on the hypothesis, indeed, one must 'suspend' it.
14. Stein, op. cit., pp. 357–9.
15. Even daltonism does not prove that certain systems are, and are alone in being, entrusted with 'seeing' red and green, since a colour-blind person manages to distinguish red if a large area in that colour is put before him, or if the presentation of the colour is made to last a long time. Id. ibid., p. 365.
16. Weizsäcker, quoted by Stein, ibid., p. 364.
17. Weizsäcker, quoted by Stein, ibid., p. 354.
18. On all these points, see *La Structure du Comportement*, in particular pp. 52 and ff., 65 and ff.
19. Gelb, *Die Farbenkonstanz der Sehdinge*, p. 595.
20. 'The sensations are certainly artificial products, but not arbitrary ones, they are the last component wholes into which the natural structures can be decomposed by the 'analytical attitude.' Seen from this point of view, they contribute to the knowledge of structures, and consequently the results of the study of sensations, correctly interpreted, are an important element in the psychology of perception.' Koffka, *Psychologie*, p. 548.
21. Cf. Guillaume, *L'Objectivité en Psychologie*.
22. Cf. *La Structure du Comportement*, chap. III.
23. Koffka, *Psychologie*, pp. 530 and 549.
24. M. Scheler, *Die Wissenformen und die Gesellschaft*, p. 412.
25. Ibid., p. 397. 'Man approaches ideal and exact images better than the animal, the adult better than

the child, men better than women, the individual better than the member of a group, the man who thinks historically and systematically better than the man impelled by tradition, 'imprisoned' in it and incapable of objectivizing, by building up recollection, the environment in which he is involved, of localizing it in time and possessing it by setting it away from himself in a past context.'

26. Hering, Jaensch.
27. Scheler, *Die Wissenformen und die Gesellachaft*, p. 412.
28. Cf. Wertheimer, *Über das Denken der Naturvölker*, in *Drei Abhandlungen zur Gestalttheorie*.

55 The Intentionality of Sensation: A Grammatical Feature

G. E. M. Anscombe

I. Intentional Objects

Berkeley calls 'colours with their variations and different proportions of light and shade' the 'proper' and also the 'immediate' objects of sight.[1] The first at any rate long seemed obvious to everyone, both before Berkeley and since his time. But Berkeley's whole view is now in some disrepute. Sense-data, a thoroughly Berkeleyan conception given that name by Russell, have become objects of ridicule and contempt among many present-day philosophers.

That word 'object' which comes in the phrase 'object of sight' has suffered a certain reversal of meaning in the history of philosophy, and so has the connected word 'subject,' though the two reversals aren't historically connected. The subject used to be what the proposition, say, is about: the thing itself as it is in reality—unprocessed by being conceived, as we might say (in case there is some sort of processing there); objects on the other hand were formerly always objects *of*———. Objects of desire, objects of thought, are not objects in one common modern sense, not individual things, such as *the objects found in the accused man's pockets*.

I might illustrate the double reversal by a true sentence constructed to accord with the *old* meanings: subjectively there must be some definite number of leaves on a spray that I see, but objectively there need not: that is, there need not be some number such that I *see* that number of leaves on the spray.

When Descartes said that the cause of an idea must have at least as much formal reality as the idea had objective reality, he meant that the cause must have at least as much to it as what the idea was of would have, if what the idea was of actually existed. The '*realitas objectiva*' of an idea thus meant what we should call its 'content'—namely what it is of, but considered as belonging purely to the idea. 'What a picture is of' can easily be seen to have two meanings: what served as a model, what the picture was taken from—and what is to be seen in the picture itself, which may not even have had an original.

Thus formerly if something was called an object that would have raised the question 'object of what?' It is hardly possible to use the word 'object' in this way nowadays unless it actually occurs in such a phrase as 'object of desire' or 'object of thought.' Suppose somebody says that the object of desire, or desired object, need not exist, and so there need not be any object which one desires. He is obviously switching from one use of the word 'object' to another. If, however, we speak of objects of sight, or seen objects, it will usually be assumed that 'objects' has the more modern sense: these will be objects, things, entities, which one sees. Now to prevent confusion I will introduce the phrase 'intentional object' to mean 'object' in the older sense which still occurs in 'object of desire.'

'Intentional' in these contexts is often spelt with an *s*. This was an idea of Sir William Hamilton's; he wanted to turn the old logical word 'intention' into one that looked more like 'extension.' I prefer to keep the older spelling with two *t*s. For the word is the same as the one in common use in connection with action.

Excerpted from Ronald J. Butler, ed., *Analytic Philosophy* (Blackwell, 1965). Reprinted with permission.

The concept of intention which we use there of course occurs also in connection with *saying*. That makes the bridge to the logician's use.

There are three salient things about intention which are relevant for my subject. First, not any true description of what you do describes it as the action you intended: only under certain of its descriptions will it be intentional. ('Do you mean to be using that pen?'—'Why, what about this pen?'—'It's Smith's pen.'—'Oh Lord, no!') Second, the descriptions under which you intend what you do can be vague, indeterminate. (You mean to put the book down on the table all right, and you do so, but you do not mean to put it down anywhere in particular on the table—though you do put it down somewhere in particular.) Third, descriptions under which you intend to do what you do may not come true, as when you make a slip of the tongue or pen. You act, but your intended act does not happen.

Intentionality, whose name is taken from intention and expresses these characteristics of the concept *intention*, is found also in connection with many other concepts. I shall argue that among these are concepts of sensation. Like many concepts marked by intentionality, though unlike intention itself, these are expressed by verbs commonly taking direct objects. I shall speak of intentional verbs, taking intentional objects. I have mentioned the history of the word 'object' to forestall any impression that 'an intentional object' means 'an intentional entity.'

Obvious examples of intentional verbs are 'to think of,' 'to worship,' 'to shoot at.' (The verb 'to intend' comes by metaphor from the last—'*intendere arcum in*,' leading to '*intendere animum in*.') Where we have such a verb taking an object, features analogous to the three features of intentionalness in action relate to some descriptions occurring as object-phrases after the verb.

The possible non-existence of the object, which is the analogue of the possible non-occurrence of the *intended* action, is what has excited most attention about this sort of verb. 'Thinking of' is a verb for which the topic of the non-existent object is full of traps and temptations; 'worshipping' is less dangerous and may help us to keep our heads. Consider the expression 'object of thought.' If I am thinking of Winston Churchill then he is the object of my thought. This is like 'What is the object of these people's worship?' Answer: 'The moon.' But now suppose the object of my thought is Mr. Pickwick, or a unicorn; and the object of my worship is Zeus, or unicorns. With the proper names I named no man and no god, since they name a fictitious man and a false god. Moreover Mr. Pickwick and Zeus are nothing but a fictitious man and a false god (contrast the moon, which, though a false god, is a perfectly good heavenly body). All the same it is clear that 'The Greeks worshipped Zeus' is true. Thus '*X* worshipped ———' and '*X* thought of ———' are not to be assimilated to '*X* bit ———.' For, supposing '*X*' to be the name of a real person, the name of something real has to be put in the blank space in '*X* bit ———' if the completed sentence is to have so much as a chance of being true. Whereas in '*X* worshipped ———' and '*X* thought of ———' that is not so.

This fact is readily obscured for us because with '*X* thought of ———' the more frequent filling-in of the blank is a name or description of something real; for when the blank is filled in so in a true sentence, it is the real thing itself, not some intermediary, that *X* thought of. This makes it look as if the reality of the object mattered, as it does for biting. Nevertheless, it is obvious that vacuous names can complete such sentence-frames. So perhaps they stand in such frames for something with a *sort* of reality. That is the hazy state of mind one may be in about the matter.

A not very happy move to clarify it is to say, 'Well, *X* had his idea of Zeus, or unicorns, or Mr. Pickwick, and that gives you the object you want.' This is an unhappy move on several counts. First, it makes it seem that the *idea* is what *X* was worshipping or thinking of. Second, the mere fact of real existence (is this now beginning to be opposed to existence of some other kind?) can't make so very much difference to the analysis of a sentence like '*X* thought of ———.' So if the idea is to be brought in when the object doesn't exist, then equally it should be brought in when the object does exist. Yet one is thinking, surely, of Winston Churchill, not of the idea of him, and just that fact started us off. When one reads Locke, one wants to protest: 'The mind is not employed about ideas, but about things—unless ideas are what we happen to be thinking about.' Whatever purpose is served by introducing ideas, by saying, 'Well, they had an idea of Zeus,' we cannot say that the idea is the object of thought, or worship. It will not be right to say *X* worshipped an idea. It is rather that the subject's having an idea is what is needed to give the proposition a chance of being true. This may seem helpful for

'worshipping,' but not for 'thinking of'; 'thinking of' and 'having an idea of' are too similar; if the one is problematic, then so is the other.

Let us concentrate on the fact that many propositions containing intentional verbs are true, and let us not be hypnotized by the possible non-existence of the object. There are other features too: non-substitutability of different descriptions of the object, where it does exist; and possible indeterminacy of the object. In fact all three features are connected. I can think of a man without thinking of a man of any particular height; I cannot hit a man without hitting a man of some particular height, because there is no such thing as a man of no particular height. And the possibility of this indeterminacy makes it possible that when I am thinking of a particular man, not every true description of him is one under which I am thinking of him.

I will now define an intentional verb as a verb taking an intentional object; intentional objects are the sub-class of direct objects characterized by these three connected features. By this definition, 'to believe' and 'to intend' are not themselves intentional verbs, which may seem paradoxical. But, say, 'to believe—to be a scoundrel' will accord with the definition, so that it is not so paradoxical as to leave out belief and intention altogether.

But now comes a question: ought we really to say that the intentional object is a bit of language, or may we speak as if it were what the bit of language stands for? As grammarians and linguists use the words nowadays 'direct object' and 'indirect object' stand for parts of sentences. So if I call intentional objects a sub-class of direct objects, that may seem already to determine that an intentional object is a bit of language.

However, the matter is not so easily settled. Of course I do not want to oppose the practice of grammarians. But it is clear that the concept of a direct object—and hence the identification of the sentence-part now called the direct object—is learned somewhat as follows: the teacher takes a sentence, say 'John sent Mary a book' and says: 'What did John send Mary?' Getting the answer 'A book' he says: 'That's the direct object.' Now the question does not really suppose, and the pupil, if he goes along with the teacher, does not take it, that any particular people, of whom the sentence is true, are in question, and so we may say that when the teaching is successful the question is understood as equivalent to 'What does the sentence 'John sent Mary a book' say John sent Mary?' The grammatical concept of a direct object is acquired by one who can answer any such question. The correct answer to such a question gives (in older usage) or itself is (in more recent usage) the direct object. Now suppose that someone were to ask: 'What is communicated to us by the phrase that we get in a correct answer? Is the phrase being used or mentioned?' It is clear that nothing is settled about *this* question by a choice whether to say, following older usage, that the phrase *gives* the direct object or, following more modern usage, that 'direct object' is a name for a sentence-part.

I propose—for a purpose which will appear—to adopt the older usage. Then the question 'What is the direct object of the verb in this sentence?' is the same as 'What does the sentence say John sent Mary?' and the question 'What does the phrase which is the answer to that question communicate to us, i.e. is it being used or mentioned?' can be asked in the form 'Is the direct object a bit of language or rather what the bit of language stands for?'—and *this* is now not a mere question of terminology, but a substantive-seeming question of curious perplexity. For someone pondering it may argue as follows: It won't do to say that in this example a book is the direct object. For if we say that we can be asked: 'Which book?'; but the sentence isn't being considered as true, and there is no answer to the question 'Which book?' except 'No book'; and yet without doubt the verb has a direct object, given by the answer 'A book.' So it must be *wrong*, and not just a matter of terminology, to say that the grammatical phrase 'direct object' stands for, not a bit of language, but rather what the bit of language stands for. And, if intentional objects are a sub-class of direct objects, the phrase 'intentional object' too will stand for a bit of language rather than what the language stands for; we are evidently not going to have to plunge into the bog made by the fact that in the most important and straightforward sense the phrase giving the intentional object may stand for nothing.

But wait—in that case *must* we not say, 'the phrase which *is* the intentional object' rather than 'the phrase giving the intentional object'? This is indeed a difficulty. For the intentional object is told in answer to a question 'What?' But the answer to 'What do they worship?' cannot be that they worship a phrase any more than that they worship an idea. A similar point holds, of course, for direct (and indirect) objects in general.

It may be argued that this is no argument.[2] Perhaps we cannot say 'What John is said to have sent is a phrase.' But then no more can we say 'What John is said to have sent is a direct object'—for the sentence did not say John sent Mary a direct object.

What this shows is that there is a way of taking 'The direct object is not a direct object' which makes this true; namely, by assimilating this sentence to 'The direct object is not a girl.' (One could imagine explaining to a child: 'The girl isn't the direct object, but the *book* that John sent.')

Frege's conclusion 'The concept horse is not a concept' was based on the same sort of trouble about different uses of expressions. What '*cheval*' stands for is a concept, and what '*cheval*' stands for is a horse; these premisses do not, however, yield the result that if Bucephalus is a horse he is a concept. Similarly, what John is said to have sent Mary is a book, and what John is said to have sent Mary is a direct object; these premisses do not yield the result that if John gave Mary a book, he gave her a direct object.

Frege eventually proposed to deal with the trouble by stipulating that such a phrase as 'What 'cheval' stands for' should *only* be used predicatively. A parallel stipulation in our case: 'What John is said to have sent Mary is . . .' may only be completed with such expressions as could fill the blank in 'John sent Mary. . . .'

The stipulation, while harmless, would be based on failure of ear for the different use of the phrase 'What John is said to have sent Mary' in the explanation 'What John is said to have sent Mary is the direct object of the sentence.' But an ear for a different use cannot be dispensed with, as the further course of the argument shows.

The argument began with stating reasons why a direct object can't be something that the direct-object phrase stands for. Yet one can, one correctly does, say 'A book' in answer to the question 'What does the sentence 'John sent Mary a book' say John sent Mary?' which asks the same thing as 'What is the direct object in that sentence?' Nevertheless the way the phrase 'a book' is being used is such that one can't sensibly ask 'Which book?'

We must conclude of 'objects' (direct, indirect and likewise intentional) that the object is neither the phrase nor what the phrase stands for. What then is it? The question is based on a mistake, namely that an explanatory answer running say 'An intentional (direct, indirect) object is such-and-such' is possible and requisite. But this need not be so. Indeed the only reasonable candidates to be answers are the ones we have failed. But what is the actual use of the term? Given a sentence in which a verb takes an object, one procedure for replying to the question: 'What is the object in this sentence?' is to recite the object phrase.

If putting the object phrase in quotes implies that the object—i.e. what John is said to have sent Mary, what the Greeks worshipped—is a piece of language, that is wrong; if its not being in quotes implies that something referred to by the object phrase is the object, that is wrong too. To avoid the latter suggestion one might insist on putting in quotes; to avoid the former one might want to leave them out. One is inclined to invent a special sort of quotes; but the question is how the phrase within such new quotes would function—and if we understand that, we don't need a new sign. So ends the argument.

To repeat, I am not opposing the practice of grammarians and linguists for whom the expression 'direct object' is defined as an expression for a phrase; they use that as I use the expression 'direct-object phrase.' But, as I have argued, the question 'What does the sentence say John gave?' is fundamental for understanding either 'direct object' or 'direct-object phrase' as I am using those expressions; and hence for understanding 'direct object' when it is used for a phrase. And though the question is answered (like many questions) by uttering a phrase—in this case 'a book'—the phrase has *a special use* in answer to that question 'What does the sentence say John gave?' *It* can name neither a piece of language, nor anything that the piece of language names or otherwise relates to, nor indeed anything else. The interest of the question and answer is the rather special interest of getting grammatical understanding. Grammatical understanding and grammatical concepts, even the most familiar ones like sentence, verb, noun, are not so straightforward and down-to-earth a matter of plain physical realities as I believe people sometimes suppose. The concept of a noun, for example, is far less of a physical concept than that of a coin; for someone might be trained to recognize coins with fair success though he knew nothing of money, but no one could be trained to recognize nouns without a great familiarity with language; and yet the concept of a noun is not one which he will automatically have through that familiarity, as he will have that of a coin if he operates with coined money. Indeed the explanations of grammatical terms are only hints at

what is really grasped from examples. Thus no one should think that by merely adopting the usage of modern grammarians, for whom the direct object is a word or words, he has avoided handling difficult concepts and remained in a plain man's world of plain thing.

> 'The direct object is what John sent' (= 'what the sentence says John sent').
> 'The intentional object is what *X* was thinking of.'

These two sentences are parallel. It is for the sake of parallelism that we opted for the old-fashioned usage of 'direct object.' For even in that usage, no one will be tempted to think that direct objects as such are a special type of entity. Just this temptation exists very strongly for objects of thought and sensation; that is, for intentional objects, which appear as entities under the names 'idea' and 'impression.'

It may be objected: the context 'The sentence says John sent Mary ———' is itself intentional. How, then, can my considerations about direct objects throw light on intentional objects? Fully spelled out they are themselves merely examples of sentences whose objects are intentional objects.[3]

The answer is that what is said in the objection is true. But these examples, where we talk about direct objects, are harmless and profitable because certain sorts of suggestion about direct objects are patent nonsense. For example no one would think that if a sentence says John sent Mary a book, what it immediately and directly says he sent her was a direct object, and only in some indirect fashion, via this immediate object, does it say he sent her a book. I want, that is, to use a comparison with patent nonsense about direct objects in order to expose as latent nonsense of just the same kind some very persuasive views about ideas and impression. Not that ideas and impressions are to be excluded from consideration; but as they enter into epistemology they will be rightly regarded as grammatical notions, whose role is readily misunderstood. And 'grammatical' is here being used in its ordinary sense.

We must now ask: does any phrase that gives the direct object of an intentional verb in a sentence necessarily give an intentional object? No. Consider: 'These people worship Ombola; that is to say, they worship a mere hunk of wood.' (cf. 'They worship sticks and stones.') Or 'They worship the sun, that is, they worship what is nothing but a great mass of frightfully hot stuff.' The worshippers themselves will not acknowledge the descriptions. Their idol is for them a divinized piece of wood, one that is somehow also a god; and similarly for the sun.

An intentional object is given by a word or phrase which gives a *description under which.*

It will help if we consider shooting at, aiming. A man aims at a stag; but the thing he took for a stag was his father, and he shoots his father. A witness reports: 'He aimed at his father.' Now this is ambiguous. In the sense in which given the situation as we have described it, this report is true, the phrase 'his father' does not give an intentional object. Let us introduce the term 'material object': 'his father' gives, we shall say, the *material* object of the verb in the sentence 'He aimed at his father' in the sense in which this was true. Not because he hit his father—he might after all merely have gone wide of the mark. But because the thing he took for a stag actually was his father. We can ask what he was doing—what he was aiming at—*in that* he was aiming at a stag: this is to ask for another description '*X*' such that in 'He was aiming at *X*' we still have an intentional object, but the description '*X*' gives us something that exists in the situation. For example, he was aiming at that dark patch against the foliage. The dark patch against the foliage was in fact his father's hat with his father's head in it.

Thus, the given intentional object (the stag) being nonexistent in the situation, we looked for another intentional object until we found one that did exist. Then the phrase giving that intentional object, and any other true description of the existent thing in question, gives the *material* object of 'He aimed at. . . .'

Does this account depend on the report's being true? No; but if the witness lies or is quite mistaken, all the same he can be questioned about what his report meant. Does he mean the phrase 'his father' to give the intentional, or only the material, object? If only the material object, what does he mean by 'He aimed at . . .'? That you could see that the man was taking aim, and where his target lay? There might not be true answers to these questions, but the witness has got to pretend there are or be confounded.

And now, for greater ease of expression, I will speak, as is natural, of the material and intentional objects of aiming, of worshipping, of thinking. This should always be interpretable in terms of the verbs and their objects.

There need not be a material object of aiming. If a man were totally hallucinated, and, shooting

at something in his hallucinatory scene, hit his father, that would not make his father the *material* object of his aiming. Similarly, if there is no description, still giving the intentional object of worship, which describes anything actual, the worshippers, materially speaking, worship a nothing, something that does not exist.

Not that it will then do to say 'They worship nothing,' but only: 'What they worship is nothing.' For 'They worship nothing' would imply that no sentence 'They worship such-and-such' will be true; and in the case supposed some such sentence is true.

Questions about the identity of an intentional object, when this cannot be reduced to the identity of a *material* object, are obviously of some interest. How do we decide that two people or peoples worship or do not worship the same god? Again, when a proper name is obscure and remote in its historical reference, like 'Arthur,' the question may arise whether two people are thinking of the same man—if they have different, incompatible, pictures of him.

But I perceive that my saying 'when this cannot be reduced to the identity of a *material* object' may mislead: for by *material* objects I do not mean what are now called 'material objects'—tables, planets, lumps of butter and so on. To give a clear instance: a debt of five dollars is not a material object in this latter sense; but given that someone had contracted such a debt, my thought 'that debt of five dollars' would have as its material object something described and indicated by the phrase giving the intentional object of my thought. When it is beyond question that the phrase giving an intentional object does describe and indicate a material object in this sense, then the question as to the identity of the intentional object reduces to the question as to the identity of the material object. Are we referring to the same debt? That is, perhaps, not too difficult to establish. But when either there is no real debt or it is very obscure whether there is, the case is altered.

The fact that we can use the concept of identity in connection with intentional objects should not lead us to think there is any sense in questions as to the kind of existence—the ontological status—of intentional objects as such. All such questions are nonsensical. Once more we can clear our heads by thinking of direct objects. The answer to 'What is the direct object in 'John sent Mary a book'?' is 'A book.' This is the right answer as much when the sentence is false as when it is true, and also when it is only made up, as it is in this case, to illustrate a point. It is evident nonsense to ask about the mode of existence or ontological status of the direct object as such: or to ask what kind of thing *a book* is, as it is thought of in answer to the question about the direct object.

II. Sensation

In the philosophy of sense-perception there are two opposing positions. One says that what we are immediately aware of in sensation is sense-impressions, called 'ideas' by Berkeley and 'sense-data' by Russell. The other, taken up nowadays by 'ordinary language' philosophy, says that on the contrary we at any rate *see* objects (in the *wide* modern sense which would include, e.g. shadows) without any such intermediaries. It is usually part of this position to insist that I can't see (or, perhaps, feel, hear, taste or smell) something that is not here, any more than I can hit something that is not there: I can only *think* I see (etc.) something if it isn't there, or only in some extended usage of 'see' do I see what isn't there. I shall say most about seeing, as most people do in discussing this topic. The other verbs are for good reasons (which aren't very relevant to my topic) often treated rather differently, especially by ordinary language philosophy.

I wish to say that both these positions are wrong; that both misunderstand verbs of sense-perception, because these verbs are intentional or essentially have an intentional aspect. The first position misconstrues intentional objects as material objects of sensation; the other allows only *material* objects of sensation; or at any rate does not allow for a description of what is seen which is e.g. neutral as between its being a real spot (a stain) or an after-image, giving only the content of an experience of seeing concerning which one does not yet know whether one is seeing a real spot or an after-image.[4]

To see the intentionality of sensation it is only necessary to look at a few examples which bring it out.

(1) 'When you screw up your eyes looking at a light, you see rays shooting out from it.'
(2) 'I see the print very blurred: is it blurred, or is it my eyes?'
(3) 'Move these handles until you see the bird in the nest.' (Squint-testing apparatus; the bird and the nest are on separate cards.)
(4) 'I see six buttons on that man's coat, I merely see a lot of snowflakes framed by this window-frame—no definite number.'

(5) '. . . a mirage. An approaching pedestrian may have no feet (they are replaced by a bit of sky).'[5]

(6) 'With this hearing aid, when you talk I hear some screeching noises; no low tones and the consonants are very indistinct.'

(7) 'I hear a ringing in my ears.'

(8) 'I heard a tremendous roaring noise outside, and wondered with alarm for a moment what great machine or floodwater could be making it. And then I realized that it was only my little dog snoring close at hand.'[6]

(9) 'Do you know how a taste can sometimes be quite indeterminate until you know what you are eating?'

(10) 'I keep on smelling the smell of burning rubber when, as I find out, there is no such thing.'

Someone who wishes to say that the verbs of sense are used right in normal cases *only* with real things as objects, and even with real things correctly characterized, may say that these are exceptional uses. Either the context (eye-testing apparatus) or what is said, with the tone of voice and special emphasis appropriate to it, shows this. There was presumably a definite number of snowflakes falling so as to be seen from a certain position, and that was the number seen; only the subject did not know how many there were, was not able to tell by looking as he could tell the number of buttons on the coat. He expressed this by saying he did not *see* a definite number of snowflakes; but this is an odd use of 'see,' different from the more normal use we get in the following example:

(11) 'I saw someone in the study just now.' 'Nonsense! You can't have, because there isn't anyone there.' 'Well, I wonder what I saw, then.'

Now this may be; on the other hand the oculist testing the degree of a squint does not have to teach a new use of 'see' or of 'I see a (picture of a) bird in a nest' before he can ask 'Do you see the bird in the nest?'—the bird-picture and the nest-picture being in fact spatially separated. To call such a use 'new' simply means that some difference between it and what is being called the old use strikes us as important.

There is indeed an important difference; though it is wrong to regard the uses which it marks as, so to speak, *deviant*, for our concepts of sensation are built up by our having *all* these uses. The difference we are attending to is that in these cases, object phrases are used giving objects which are, wholly or in part, merely intentional. This comes out in two features: neither possible non-existence (in the situation), nor indeterminacy, of the object is any objection to the truth of what is said.

Now 'ordinary language' views and 'sense-datum' views make the same mistake, that of failing to recognize the intentionality of sensation, though they take opposite positions in consequence. This failure comes out clearly on the part of an ordinary-language philosopher if he insists that what I say I see must really be there if I am not lying, mistaken, or using language in a 'queer,' extended (and therefore discountable) way.

The Berkeleyan sense-datum philosopher makes the same mistake in his insistence that, e.g., one sees visual impressions, visual data. I would say that such a philosopher makes an incorrect inference from the truth of the grammatical statement that the intentional object, the impression, the visual object, is what you see. He takes the expression 'what you see' materially. 'The visual impression is what you see,' which is a proposition like 'The direct object is what he sent,' is misconstrued so as to lead to 'You see an impression,' as the other never would be misconstrued so as to lead to 'He sent her a direct object.'

This is a more interesting and permanently tempting mistake than the other, whose appeal is merely that of a common-sense revolt against a Berkeleyan type of view. But both doctrines have a great deal of point. To take the 'ordinary language' doctrine:

First, what I shall call the material use of verbs of sense exists. The material use of 'see' is a use which demands a *material* object of the verb. 'You can't have seen a unicorn, unicorns don't exist.' 'You can't have seen a lion, there wasn't any lion there to see.' These uses are quite commonplace. It is not merely that the object-phrase is taken materially—as we have seen, that may be the case with an intentional verb without reflecting on its intentionality. Here the verb 'to see' is not allowed to take a *merely* intentional object; non-existence of the object (absolutely, or in the situation) is an objection to the truth of the sentence. We see the double use of the verb 'see' by contrasting it with 'worship.' No one would ever say: 'They cannot have worshipped unicorns, because there are no such things.'

Second, the words giving the object of a verb of sense are necessarily most often intended as

giving *material* objects of sense: for this is their primary application. To see this, consider the following. Suppose a bright red plastic toy elephant looks greyish-brown to me in a certain light. Only if I do not know that the greyish-brown colour is mere appearance do I say without any special context (e.g. that of describing impressions), or apology, or humour: 'I see a greyish-brown plastic toy elephant.' This is because we understand the description-of-an-appearance 'greyish-brown' by understanding the description 'greyish-brown': this describes what the appearance is of. To do that, it must in the first instance be a description of such a thing as it would be true of (for the appearance is an appearance of that)—really, and not merely in appearance: this will be its primary application. But, being a description of a sensible property, it must also in its primary application enter into the object phrases for the appropriate verbs of sense, since we get to know sensible properties by the appropriate senses.

Further, we ought to say, not: 'Being red is looking red in normal light to the normal-sighted,' but rather 'Looking red is looking as a thing that *is* red looks in normal light to the normal-sighted.' For if we ought rather to say the first, then how do we understand 'looking red'? Not by understanding 'red' and 'looking.' It would have to be explained as a simple idea; and so would looking any other colour. It may be replied: These all are simple ideas; 'looking yellow' and 'looking red' are the *right* expressions for what you show someone when you show him yellow and red, for he will only learn 'yellow' and 'red' from the examples if they look yellow and look red; so it is *looking-yellow* and *looking-red* that he really gets hold of and has been introduced to, even though you *say* you are explaining 'yellow' and 'red.' This would come to saying that in strictness 'looking' should be part of every colour word in reports of perception: it will then cease to perform the actual function of the word 'looking.' It was plausible to say: Only if it looks red to him will he learn what is meant; but wrong to infer: What he then grasps as the correlate of the word 'red' is a red look. Even granted that he knows he is to learn the name of a colour, still it invites misunderstanding to rely on something that only *looks* red to teach him the word; if he notices that it only looks red, how natural for him to suppose that 'red' was the name of the colour that it actually *is*. If you tell him: 'It's the colour that this 'looks,' ' this presupposes that 'looks *C*' and '*C*' are originally, and not just subsequently, distinct: that, in short, 'being red' is not after all to be explained as a certain looking-red.

Again, things do not always look the same shape, colour, size and so on, but we commonly look at and describe them, saying, e.g., 'It's rectangular, black and about six foot in height,' without paying attention to how they look—indeed we might say that often things *look* to us, strike us, not as they look but as they are! (Conviction that *only* so is 'looks' used rightly was the cause of confusion to an overconfident ordinary-language philosopher on an occasion famous in Oxford: F. Cioffi brought in a glass vessel of water with a stick in it. 'Do you mean to say,' he asked, 'that this stick does not look bent?' 'No,' said the other bravely: 'It looks like a straight stick in water.' So Cioffi took it out and it *was* bent.)

So much at least there is to be said on the side of the 'ordinary-language' philosopher. But, turning to the sense-impression philosophy, how much it points out and can investigate which often gets querulously dismissed by the other side! There is such a thing as simply describing impressions, simply describing the sensible appearances that present themselves to one situated thus and thus—or to *myself*.

Second, the sense-impression philosophy will be right in its way of taking the Platonic dictum: 'He who sees must see something.' Plato compared this to 'He who thinks must think something,' and has sometimes been criticized on the ground that 'seeing' is a relation of a subject to an object in the modern sense of that last word, while thinking is different: that such-and-such is the case isn't a thing. But 'He who sees must see something' is being wrongly taken if taken as meaning: 'Whenever anyone can rightly be said to see, there must be something there, which is what he sees.' Taken in that sense, it is not true; to say it is true is to legislate against all except the material use of 'see.' The sense in which it is true is that if someone is seeing, there is some content of his visual experience. If he says he can see ('can see' is English idiom for 'is seeing') we can ask him 'What can you see?' He may say 'I don't know.' Perhaps that means that he doesn't know what the material object of his seeing is; perhaps simply that he is at a loss to make out *what* what he (in any sense) sees *looks like*. But then we can say: well, at any rate, describe what colours, what variation of light and dark you see. He may say: 'It's frightfully difficult, it all changes so fast, so many colours shifting all the time, I can't

describe it, it doesn't stay long enough'—and that's a description. But he cannot say: 'how do you mean, what I see? I only said I could see, I didn't say I could see something—there's no need of a '*what*' that I see.' That would be unintelligible.

This brings out the third point in favour of the sense-impression philosophy, which offers it some support even in its strict Berkeleyan form. The minimum description that must be possible if someone can see, will be of colours with their variations of light and darkness. One cannot say 'Colour, light and dark? No question of any such things,' in response to a *present* enquiry about what one sees.

That is to say, it is so with us. Perhaps we could imagine people whose language has no colour vocabulary, though they are sighted, i.e. they use eyes and need light to get about successfully, etc. A man of such a people, taught to read by sight, learns names of letters, could read out words which were black on white, but could not understand the words 'black' and 'white.' We'd say we do not know 'how he tells' the words, the shapes. But is that to say anything but that for us appeal to colours is used in an account of how we tell shapes? Whereas perhaps for him there is in this sense no such thing as a 'how he tells'—any more than there is for us with the colours themselves. We don't ask for a 'how we tell' it's red, as we ask for a 'how we tell' it's the word 'red' and accept as part of the answer 'by seeing these shapes, i.e. colour patches of these shapes.' We may wonder 'How could there be such recognition of a thing like the pattern of a word—*unmediated* recognition? How could it but be mediated by perception of colour?' (One of the origins of the notion of simple ideas, elements.) But although in this case we have an account of the perception of the pattern as mediated by the perception of colour, think of our recognition of human expressions. We feel that this is the *kind* of thing to be mediated, but fail in our attempts to describe the elements and their arrangements, seeing which we recognize a cheerful or ironical expression. But, one may say, optically speaking he must be being affected by light of the wavelengths belonging to the different colours. Yes—but does that show that, so to speak, the content of a colour concept is pushed into him, so that all he has to do is utter it in a name, whose use he will later make to fit with other people's in its range of application? I believe this is thought. (cf. Quine about 'square' and each man's retinal projection of a square tile.)[7] Formulated, this loses its plausibility. For one thing, the optical process does not exhibit anything to the man in whom it takes place. For another, no concept is simply given; every one involves a complicated technique of application of the word for it, which could not just be presented by an experience-content. The fact that there is no 'how we tell' about colour-recognition does not mean that training in practices—most strikingly the practices comprising that technique of application—is not as necessary for the acquisition of colour concepts as those of substances or square roots.

Pursuant to this false conception of the primitively given, Berkeley—and Russell—thought that all else in description of the seen, all besides the arrangement of colour patches in the visual field, was inference and construction. This is not acceptable. There are impressions of distance and size, for example, independent of assumptions about what a thing is. One may be utterly perplexed what a thing is just because one is seeing it as at a different distance from the right one, and hence as the wrong size. Or vice versa. I once opened my eyes and saw the black striking surface of a matchbox which was standing on one end; the other sides of the box were not visible. This was a few inches from my eye and I gazed at it in astonishment wondering what it could be. Asked to describe the impression as I remember it, I say: 'Something black and rectangular, on end, some feet away, and some feet high.' I took it for three or four feet distant, and it looked, if anything, like a thick post, but I knew there could be no such thing in my bedroom. Or I have taken a small black prayer book for a great family Bible sort of volume, judging that it lay on a footrest some feet away instead of a nearby ledge nearer eye-level. These were not judgements of distance based on identifications of things—the supposition of what thing it might be was based on an impression of size which went with a false impression of distance.

Departing, then, from Berkeley, we can note that descriptions of visual impressions can be very rich and various. There can be impressions of depth and distance and relative positions and size; of kinds of things and kinds of stuff and texture and even temperature; of facial expression and emotion and mood and thought and character; of action and movement (in the *stationary* impression) and life and death. Even within the compass of the description 'colours with their variations of light and shade' there are diverse kinds of impression.

It remains to sort out the relations between the intentional and material objects of sensation; as I have done most of the time, I will concentrate on seeing.

While there must be an intentional object of seeing, there need not always be a material object. That is to say '*X* saw *A*' where 'saw' is used materially, implies some proposition '*X* saw ———' where 'saw' is used intentionally; but the converse does not hold. This leads to the feeling that the intentional use is somehow prior to the material use. The feeling seems to run contrary to the recognition, the feeling, that for descriptions of objects of sight the material application is the prior one. Both feelings are—legitimately—satisfied by allowing that an intentional object is necessarily involved in seeing, while granting that this does not confer epistemological priority on purely intentional sentences, which indeed, in a host of the most ordinary cases of reported seeing, are never formulated or considered.

John Austin, who opposed the view that there are two senses of 'see' according as the seeing has to be veridical or not, remarked casually that there were perhaps two senses of 'object of sight.' I think it was in this connection that he contrasted 'Today I saw a man born in Jerusalem' and 'Today I saw a man shaved in Oxford'—both said in Oxford. At any rate, one says, you didn't *see* him born today; perhaps you did see someone being shaved. So the one description, while true of what you saw, in a sense does not give what you saw. A description which is true of a material object of the verb 'to see,' but which states something that absolutely or in the circumstances 'you can't have *seen*,' necessarily gives *only* a material object of seeing.

In speaking of the material object of aiming, I said that if a man aimed at that dark patch against the foliage, and that patch was his father's hat with his father's head in it, then his father was a material object of his aim; but if he aimed at some patch in a totally hallucinatory scene, and hit his father, you could not say that.

Now if we try to apply this explanation to the case of seeing we run into difficulties which reflect back on the case of aiming. But in the case considered the material object of aiming was arguably an *intentional* object of seeing. For what else—it might be asked—is a dark patch against the foliage?

This may seem to plunge us into confusion. For surely what is *only* an intentional object of seeing can't be a material object of aiming? Then when does a description give a material object of sight? One kind of case we have seen: when a description is true of what is seen, but does *not* give an intentional object. 'I see a man whose great uncle died in a lunatic asylum'—the relative clause gives an absolutely non-intentional description. 'I see a girl who has a mole between her shoulder-blades'—in the circumstances it gives a non-intentional description. For she is facing me, etc. 'You can't have *seen* that,' one says.

But why? If I can't see that, why can I see Professor Price's tomato? It has a backside that I don't see. Mr. Thompson Clarke draws our attention to the fact that a view of a tomato and a half-tomato may be exactly the same. That is so; but it is not like the fact that a view of someone with and without a mole between his shoulder blades may be exactly the same. If you look at a tomato and take only a single view, you *must* see what *might* be only a half tomato: that is what seeing a tomato is. Whereas there is a view of the mole; and no front view *is* a view of a mole between the shoulder blades. Such a mole does not stamp the front view as may approaching death or a load of troubles, and so there is no impression of it—just as there is no 'born-in-Jerusalem' look about a man.

But a material object of seeing is not necessarily given by a description of what is before my eyes when they are open and I am seeing; if I am totally hallucinated, then in no sense do I see what is before my eyes. Thus it is essential to a material object of seeing that it is given by a description which is true of *what is seen*; and we have to enquire into the significance here of this phrase '*What is seen.*'

The problem is this: there is a material object of φ-ing if there is a phrase giving an intentional object of φ-ing which is also a description of what exists in a suitable relation to the φ-er. Now this can't be a description of what exists merely by describing the intentional object of some *other* act (he aims at the dark patch that he sees); if simply describing an intentional object of φ-ing will not—as of course it will not—guarantee that we have described a material object of φ-ing, then how can it give a material object of some other verb, φ-ing?

All would be plain sailing if we could say: we have a material object of sight only if *some* intentional description is also true of what really—physically—exists. And perhaps we can say that the dark patch against the foliage is not merely an intentional object of seeing; there really is a dark object or a region of darkness there.

But this is not always the case when we see. Suppose I have defective sight: all I see is a shiny blur over there. That blur, we say, is my watch. We therefore say I see my watch, though very indistinctly; and I want to say that my watch is the material object of seeing. But I may not be able to see it as a watch; all I see is a shiny blur. But the description 'a shiny blur' is not true of anything that physically exists in the context. Supposing the father had a dark hat on, it would follow that, to mention the puzzle that perplexed Moore for so long, the dark patch against the foliage was *part of the surface of a material object* (*modern sense*); but certainly 'a blur' is no part of the surface of my watch. But it may be I have no other description of what I see than 'a shiny blur over there.' So is there any intentional description which is also a description of a material object of sight?

Yes; for even if my watch is not a blur, it is a shiny thing and it is over there. Suppose I had said: I see a roughly triangular red blur here, and some causal connection via the visual centres in the brain could have been discovered between that and the presence of my watch over there—would it have been right to say: 'What I am seeing is my watch'? I believe not.

An interesting case is that of *muscae volitantes*, as they are called. You go to the doctor and you say: 'I wonder if there is something wrong with my eyes or my brain? I see'—or perhaps you say 'I *seem* to see'—'floating specks before my eyes.' The doctor says: 'That's not very serious. They're there all right' (or: 'You see them all right')—'they are just the floating debris in the fluids of the eye. You are a bit tired and so your brain doesn't knock them out, that's all.' The things he says you see are not *out there* where you say you see them—*that* part of your intentional description is not true of anything relevant; but he does not say that what you are seeing is that debris *only* because the debris is the cause. There really are floating specks. If they caused you to see a little red devil or figure of eight, we should not say you saw them. It may be possible to think of cases where there is nothing in the intentional object that suggests a description of what is materially being seen. I doubt whether this could be so except in cases of very confused perception—how could a very definite intentional description be connected with a quite different material object of seeing? In such cases, if we are in doubt, we resort to moving the supposed material object to see if the blurred, not colour-true and misplaced image of it moves.

When you said: 'I see'—believing that the objects were quite illusory—you *intended* your description purely as an intentional one; you were giving the words 'floating specks' a secondary application. It came as a surprise to you that you would have had the right to intend the words materially. In the well-known case of H. H. Price's mescaline illusion, when without any derangement of his judgement he was able to describe what he saw—a great pile of leaves on his counterpane, which he knew not to be there—we again have a secondary application: the words 'a pile of leaves' were intended *only* as a description of an impression.

It is important to notice that very often there is no answer to the question whether people intend the word 'see' in its *material* use or not: that is, whether they are so using the word 'see' that they would have to take it back supposing that what they said they saw was not there. If they were mis-seeing something that was there, they would usually want to correct themselves, finding out 'what they really saw.' But what if the seeing were hallucinatory?

The question would be: supposing that turned out to be the case, would you claim that you mean 'see' in such a way that all you have to do is alter your intentions for the description of the object, from intending it in its *primary* application as a description of the *material* object of sight to intending it in a *secondary* application as a description of a mere *impression*?

Faced with such a question, we have in general the right to reject it, saying like Tommy Traddles: but it isn't so, you know, so we won't suppose it if you don't mind. And even if we have not this right, we generally entertain no such supposition and *therefore* are unprepared with an answer. We need not have determinately meant the word 'see' one way or the other.

We may make a similar point about 'phantom limb.' I take the part of the body where pain is felt to be the object of a transitive verb-like expression 'to feel pain in ———.' Then when there is, e.g., no foot, but *X*, not knowing this, says he feels pain in his foot, he may say he was wrong ('I did not see a lion there, for there was no lion') or he may alter his understanding of the phrase 'my foot' so that it becomes a purely intentional object of the verb-like expression. But it need not be determined in advance, in the normal case of feeling pain, whether one so intends the expression 'I feel pain in ———' as to withdraw it, or merely alters one's intentions for the description of the place of the pain, if one should learn that the place was missing.

NOTES

1. Throughout this paper single quotes are used for ordinary quotations (and so doubles for quotes within quotes) and singles I use as scare quotes.
2. This was argued to me by Mr. G. Harman, for which I am obliged to him.
3. I am indebted for this objection and the discussion of it to Professors Bernard Williams and Arthur Prior and Mr. P. T. Geach.
4. I am obliged to Professor Frank Ebersole for telling me of an experience of his which supplied this example.
5. Example from M. Luckiesh.
6. Example from W. James.
7. *Word and Object* (Cambridge, MA: 1960), p. 7.

The Limits of Self-Awareness

M. G. F. Martin

The disjunctive theory of perception claims that we should understand statements about how things appear to a perceiver to be equivalent to statements of a disjunction that either one is perceiving such and such or one is suffering an illusion (or hallucination); and that such statements are not to be viewed as introducing a report of a distinctive mental event or state common to these various disjoint situations.

When Michael Hinton first introduced the idea, he suggested that the burden of proof or disproof lay with his opponent, that what was needed was to show that our talk of how things look or appear to one to be introduces more than what he later came to call perception-illusion disjunctions:

> I do not at present see how it can be, or could be, shown that there is such a thing as (Q) [a statement which reports the occurrence of a visual experience in contrast to expressing a perception-illusion disjunction]. Consequently I do not see how it can be shown that there is such a thing as my psi-ing for these and other statements to be about; and since one surely should not make statements without being able to show that they are about something, this means that as far as I can see no such statements should be made. Perhaps I just can't see far enough, but I should like to be shown that this is so. (Hinton 1967, p. 220)

I suspect that many readers on encountering either Hinton's presentation of disjunctivism or the accounts of it available from Snowdon or McDowell, would find surprising this demand that the burden of proof for the existence of a non-disjunctive sensory experience. Surely we know what a sensory experience is in just the sense that Hinton is denying. What we don't know, the line of the thought may go, is quite what the disjunctivist is saying in its place. Doesn't the burden of proof lie, then, with the disjunctive theory of appearances: first to clarify further what it has to say, and then to offer some appropriate defence of these outlandish claims?

The aim of this paper is to offer some of that elaboration, but also in turn to explain the way in which Hinton was correct in his challenge. Properly understood, the disjunctive approach to perception is the appropriate starting point for any discussion of the nature of perceptual experience. The key to the approach is not in its appeal to paraphrasing claims about experience in disjunctive form, but is rather in an appeal to the idea of indiscriminability in explicating the claims we accept about experience. The core thought is that we grasp the idea of sense experience as such, in contrast to sense perception, through recognising that there are things that we cannot know about ourselves just through reflection on the situation we find ourselves in. As I aim to explain below, a suitable modesty about what one can know about one's experiential state is the proper starting point for theorising about sense experience in general. Any theory which moves beyond such modesty and

Excerpted from *Philosophical Studies* 120: 37–89, 2004.

makes substantive claims about the properties that sense experiences possess needs to justify this boldness.

In what follows, there are two morals that I wish to draw. The first concerns the question of the conception of sense experience in general; the second concerns the claims that the disjunctivist is committed to concerning a special sub-class of hallucinations, those brought about through the same proximate causal conditions as veridical perceptions. It is in relation to the latter that the most striking (and many will find most implausible) claim that the disjunctivist makes, that there may be sensory states whose mental nature is characterisable in nothing but epistemological terms, in terms of their unknowable difference from cases of veridical perception. But the significance of this commitment can be understood only in the light of the former claim about experience in general. Before addressing those matters, I want briefly to raise two others: the prime motivation for endorsing disjunctivism, and the question of how we are to understand its formulations as provided by Hinton, Snowdon, and McDowell.

1. The prime reason for endorsing disjunctivism is to block the rejection of a view of perception I'll label *Naïve Realism*. The Naïve Realist thinks that some at least of our sensory episodes are presentations of an experience-independent reality. When I sit here writing this, I am conscious of the various elements that make up a North London street scene. The same objects and aspects of these objects which I can attend to as part of the environment beyond me are also aspects of what I can attend to when I pursue the question, 'What is it like for me now so be staring out of the window rather than writing my paper?' Mind-independent reality can form the subject-matter of sensuous experience. In affirming this the Naïve Realist finds common ground with those views of perception which attribute to it a representational or intentional content and seek to explain its phenomenal character in terms of that content—*Intentional Theories of Perception*. For it is common to support such theories by pointing out that our sense experience is transparent—that experientially we are presented with a mind-independent realm and not simply some array of mind-dependent qualities or entities whose existence depends on this awareness.[1]

The Naïve Realist, however, claims that our sense experience of the world is, at least in part, non-representational. Some of the objects of perception—the concrete individuals, their properties, the events these partake in—are constituents of the experience. No experience like this, no experience of fundamentally the same kind, could have occurred had no appropriate candidate for awareness existed. In this, sense perception contrasts with imagining and thought. For one can certainly imagine objects in their absence, so the mind's direction on an object does not require that it actually exist when one imagines. The same is true, arguably, of thought—we think of objects which in fact do not exist as well thinking of the existent. The Naïve Realist insists that sensing is not like this, and in that respect the Naïve Realist finds common ground with the Sense-Datum tradition, or what more broadly I will label *Subjectivism*. For Subjectivists have long insisted that what is distinctive of sensing as opposed to thinking is that one really cannot sense in the absence of an object of sensing.[2]

Whatever its other merits, Naïve Realism is inconsistent with two assumptions which are common to much of the philosophical discussion of perception. The first of these is *Experiential Naturalism*: our sense experiences, like other events or states within the natural world, are subject to the causal order, and in this case are thereby subject just to broadly physical causes (i.e. including neurophysiological causes and conditions) and psychological causes (if these are disjoint from physical causes). One can manipulate the world so as to induce a hallucination in someone, for example, by suitable stimulation of their sensory cortices and possible manipulation of their psychological condition. One does not, in addition, have to invoke any further influence over other superluminary entities, something neither physical nor mental, in order to bring about the experience. The second assumption is *The Common Kind Assumption*: whatever kind of mental event occurs when one is veridically perceiving some scene, such as the street scene outside my window, that kind of event can occur whether or not one is perceiving. One may hold to this assumption for different reasons—it is tempting to suppose that it is obvious just because by 'sensory experience' we mean to pick out that event for which there is something it is like for the subject when they perceive or hallucinate, or whatever. And we are, of course, aware that from the subject's point of view there may seem to be no difference at all between a case of hallucination and one of perception. So the event in question must be of the same phenomenal kind as the kind of veridical perception it

matches. One may also eschew phenomenological evidence for the commitment in favour of an appeal to causal considerations—that reflection on how we can bring about perceptions and hallucinations should lead us to suppose that the immediate effects of appropriate brain stimulation, the experiences caused, must be the same whether or not a perception or a hallucination is brought about. Either way, the assumption is that when we are thinking of the mental or subjective aspect of perception we pick out a kind of event or state which is common to cases of perception and hallucination.

Naïve Realism together with these two assumptions leads to contradiction. For first, assume that we have some event which is as the Naïve Realist supposes a perception can be: it is an awareness of some lavender bush which exists independent of one's current awareness of it. By the Common Kind Assumption, whatever kind of experience that is, just such an experience could have occurred were one merely hallucinating. By Experiential Naturalism, we know that there are sufficient appropriate physical and psychological causes of it. If the hallucinatory experience were relational in the manner that the Naïve Realist supposes the perception of the bush to be, then the causes sufficient to bring about the hallucination must also have been sufficient for some appropriate object to be present in the experience. By our assumption about the causes, this is done without assuming any extra causal correlations between the causes of the experience and any non-physical object of awareness. Hence the bringing about of the experience must have been sufficient for the existence of its object—that is, the experience is of a kind sufficient for the existence of its object. If the experience alone is constitutively sufficient for this object of awareness in the case of hallucination, then the object in this case is not merely non-physical but dependent for its existence on the occurrence of this experience.[3,4]

Since the experience which occurs when one is hallucinating is of just the sort that occurs when one is veridically perceiving, the experience one has when one is veridically perceiving is by itself sufficient to constitute the existence of its object of awareness. According to the Naïve Realist, the object of awareness is experience-independent, yet in this case we are to suppose that there is in addition an experience-dependent object sufficient to account for the nature of the experience. So, contrary to the Naïve Realist's starting assumption, if the hallucinatory experience is a relation to an object of awareness, it is to a mind-dependent one, and hence the perception is a relation to a mind-dependent object, not the mind-independent object that the Naïve Realist hypothesises.

What if one assumes instead that the hallucinatory experience is not the awareness of anything at all? From the subject's perspective it may seem as if there is a table there before him or her, but in reality there is nothing for them to be standing in such a relation of awareness. We have to describe the situation as if there is such an object—we say that the subject 'sees' a bush, or it is 'as if' there is an apparent bush. In doing so, though, we do not really indicate any acceptance of ontological commitment; rather we treat the experience rather as having an 'intentional object.'[5] To make this move is to assume that the experience of the kind that the subject has when hallucinating does not need to have any objects of awareness as constituents of the experience—some experiences we treat as if they are the presentations of such objects, but they don't need any such objects to exist in order for them to occur. This position may seem to have the ontological advantage of avoiding any commitment to mind-dependent entities, and for that reason has often been preferred. But despite this advantage, it offers no respite from the argument we are now considering. Since the experience in question, the hallucination, is of just the same kind as the veridical perception, then the same holds of the veridical perception as of the hallucination. That is, the veridical perception does not have the objects of perception as constituents, despite the Naïve Realist's claims to the contrary.

So, Experiential Naturalism imposes certain constraints on what can be true of hallucinatory experiences. Such experiences either can have only experience-dependent objects, or not be relations to objects at all. By the Common Kind Assumption, whatever is true of the kind of experience that one has when one is hallucinating, the same must be true of the kind of experience one has when perceiving. So either one's experience when veridically perceiving is of some mind-dependent object, or the experience is not essentially a relation to any object at all.

Hence, Naïve Realism is inconsistent with these two assumptions. One way of reading the history of philosophy of perception is to see it in terms of a conflict between Naïve Realism and the kind of commitments reflected in these two assumptions.[6] Sense-datum theories hold on to one aspect of Naïve Realism, that experience is

a relation between the subject and some object of awareness, yet reject the thought that such objects can be the objects in the world around us. Intentional theories of perception are often moved by the thought that one should hold on to the other aspect of Naïve Realism, that one is related to the world around us through perceptual consciousness, but thereby give up the element of Naïve Realism, that such awareness is genuinely a relation to such objects.

The motivation for disjunctivism, I suggest, is a desire to hold on to Naïve Realism. For reasons expanded on elsewhere, I suggest that we should think of Naïve Realism as the best articulation of how our experiences strike us as being to introspective reflection on them.[7] It is common to complain against sense-datum theories that they deny that we have genuine awareness of objects in the world around us, to complain that they introduce a veil of perception. The best sense one can make of this complaint is really that sense-datum theories are forced to say that the real nature of our sensory experience is not how it strikes us as being. But if Naïve Realism is the correct description of how our sensory experience strikes us, then an intentional theory of perception is no less revisionary than a sense-datum account. To hold on to our Naïve view of experience, though, we need to reject one of the starting assumptions: either Experiential Naturalism or the Common Kind Assumption.

Experiential Naturalism was implicitly rejected by the early sense-datum theorists who were sceptical of the completeness or unity of the physical world and open to positing the existence of many strange items.[8] Such a rejection may not be *a priori* incoherent but it comes at high cost. So too does a rejection of this argument through embracing transcendental idealism, as Merleau-Ponty suggests, and as Jerry Valberg more recently has recommended.[9] If we do not think of our experience of the world as itself being a part of the world, then we need not conceive it as having causal antecedents within the world—we then need not think of how such events can otherwise be brought about.

The disjunctivist response, however, remains committed to the broad empirical assumptions and methodological presuppositions which lead one to endorse Experiential Naturalism and hence the conclusions drawn from it about the nature of our experiences. It seeks to resist the rejection of Naïve Realism, therefore, simply by denying the Common Kind Assumption. That is, we hold on to Naïve Realism by insisting that the fundamental kind of event that one's sensory experience which is a veridical perception of the table in front of one is a kind of event which just could not occur were one hallucinating. Even if some matching hallucination would either have to be an awareness of some mind-dependent object or of no object at all, nothing follows from that alone about the status of one's veridical perception.

2. What does the denial of the Common Kind Assumption amount to? The three disjunctivists with which I started, Hinton together with Paul Snowdon and John McDowell, offer significantly different formulations of the view. Contrast Hinton in the first quotation, with Snowdon and then McDowell:

> Even if few things are certain, it is certain that there are what I shall call perception-illusion disjunctions: sentences or statements like 'Macbeth perceives a dagger or is having that illusion,' which you can compose by adding words like '. . . or *x* is having that illusion' to a sentence which says that a particular person, *x*, perceives a thing of some particular kind. (Hinton 1973, p. 37)
>
> It looks to *S* as if there is an *F*: (there is something which looks to *S* to be *F*) *or* (it is to *S* as if there is something which looks to him (*S*) to be *F*). (Snowdon 1980–1981)
>
> . . . an appearance that such-and-such is the case can be either a mere appearance or the fact made manifest to someone . . . the object of experience in the deceptive cases is a mere appearance. But we are not to accept that in the non-deceptive cases too the object of experience is a mere appearance, and hence something that falls short of the fact itself . . . appearances are no longer conceived as intervening between the experiencing subject and the world. (McDowell 1982)

In each case the disjunctive form is specified in significantly different ways. Hinton and Snowdon focus on locutions of object perception, '*S* sees *o*,' which are commonly taken to be transparent in the object position. Hinton contrasts on either side of his disjunction the seeing of a flash of light with the having of an illusion of a flash of light. Snowdon, in contrast, treats both veridical perception and illusion as belonging on the privileged side of the disjunction, since both involve perception of an object, and keeps only hallucination to the contrasted side. In contrast to both of these, McDowell is interested in locutions of factual perception: '*S* sees/can see that *p*.' Such locutions are

typically opaque in the complement clause, and the relation between talk of object perception and fact perception is complex – not every object mentioned in a perceived fact need be an object of perception; even if some fact must be perceived concerning any object of perception, it is not clear that there is any specific fact which must have been perceived in perceiving an object. McDowell's contrast case, then, is simply that of merely apprehending the appearance of *p*, rather than properly grasping the fact.[10]

Perhaps, then, rather than speaking in terms of *the* disjunctive theory of appearances, we should recognise a cluster of approaches, all of which have in common just a negative thesis: the thesis that we should not think that perceptual experience is to be analysed as a common factor of perception and either illusion or hallucination.

However, if we characterise the approach just in this negative way, then we are also liable to be misled. For this seems to offer merely an incomplete sketch of an account which needs further supplementation. First, one needs some further gloss of the 'privileged' disjunct – the reference to perception or veridical perception. That there is some idea at the back of these theories is often implicitly understood when one reads them—these accounts are supposed, somehow or other, to defend some form of direct realism. On the other hand, the negative construal as yet does not tell us what to say about the 'underprivileged' disjunct, the one that fails the condition to be counted as perception. Surely we need to know what more to say about these cases before we know what these approaches are telling us about perceptual experience in general. Jonathan Dancy offers a clear expression of this line of thought when he suggests:

> The disjunctive account of perception really says that there are two quite different sorts of oasis-experience, which may none the less be indistinguishable to their owner. The first is the genuine article, and the second, though it is indistinguishable, has nothing in common with the first other than the fact that they are both oasis-experiences. In the standard formulation of the account, misleadingly, this is explicitly the way in which the second disjunct is characterized: we characterize it solely by saying that it is like what it is not. Presumably, however, there may be available a more direct characterization of the second disjunct, and in a totally explicit version of the theory it would indeed be characterized in that better way. The current characterization is just the sort of place-holder, showing what has to be said about the relation between the first and second disjunct. (Dancy 1995)

Yet if we take Dancy's concerns seriously and attempt to spell disjunctivism out in more detail, we encounter two further problems. According to Dancy, in the proper dress of the theory, we should surmise that the full account of perceptual experience offers a clause for the privileged case of perception and the underprivileged case of illusion or hallucination. Where other theories can hope to offer a common explanation of the phenomena that we look to perception and sensory experience to provide, the supplemented disjunctivism will need to offer two distinct accounts. As such the approach necessarily lacks the consilience of conjunctive accounts of sensory experience. Struck by this obvious thought, it is no surprise that opponents are liable to think that justification needs to be provided for the disjunctivist position rather than vice versa—Hinton's attitude is liable to seem mere complacence.

Behind this lies a deeper worry. Is there really a coherent supplementation to the disjunctive account? Suppose we do get a further specification of the kind of mental event that occurs in the non-privileged circumstances. If what marks these cases out in the first place is just that they involve the absence of perception, then one may worry that whatever fixes what they have in common with each other will apply equally to any case of perception. That is to say, the further specification of hallucination will be something which is present not only in all cases of illusion or hallucination but also in the case of perception. The disjunctivist will then be left in the unhappy position of conceding that there is a common element to all of the cases, while still insisting that there is something distinctive of perception. Now if the common element is sufficient to explain all the relevant phenomena in the various cases of illusion and hallucination, one may also worry that it must be sufficient in the case of perception as well. In that case, disjunctivism is threatened with viewing its favoured conception of perception as explanatorily redundant.

In what follows I will address both of these concerns. Hinton is right to say that the disjunctivist conception of perceptual experience in general should in fact be our default conception. And Dancy is wrong to think that the disjunctivist specification is incomplete, that we should supplement the account of experience with a non-relational gloss of what illusory or hallucinatory

experiences are. Nonetheless, the remaining worry about explanatory redundancy does pose a serious challenge to disjunctivism, as we shall see, and in the end addressing this challenge brings out the most distinctive and surprising aspect of disjunctivism: the limits to the self-awareness we can have of our own sensory states.

Properly understood, disjunctivism offers us an epistemological perspective on how we should conceive the debate about sensory experience. It helps bring out how weighty one's epistemological assumptions about the mind must be, if one is to advance beyond this epistemological stance.

3. How then should we think about sense experience? What gives us a grip on the notion? Contrast two different ways of thinking about the Cartesian story of lucid dreaming. Few of us have any problem grasping the idea of perfect hallucinations. At present, I have good reason to suppose that I am seeing a London street pretty much as it is. So I have a veridical perception of the unkempt lavender bush at the end of my road that marks the advance of late summer. Nonetheless, as far as I can tell, it seems a genuine possibility that I could have been in a situation which was not one of actually perceiving my environment for how it was but which I would not have been able to tell apart from this, my actual situation, just through introspection and reflection on my experience. Such a case would surely be a perfect hallucination of the kind of scene that I am perceiving, as things stand, for what it is.

On the first conception of experience, one that someone who endorses the Common Kind Assumption might endorse, this starting point is further elaborated so. A perceptual experience is a kind of event which has certain distinctive features $E_1 \ldots E_n$. Not only is the possession of these features necessary and sufficient for an event to be an experience, but, in addition, an event's possession of them is introspectible by the subject of the experience. When I come to recognise the possibility of perfect hallucination just like my current perception, what I do is both recognise the presence of these characteristics, $E_1 \ldots E_n$, in virtue of which this event is such an experience, and also recognise that an event's possessing these characteristics is independent of whether the event is a perception or not. So in accepting the Cartesian possibility I display a grasp of a positive piece of knowledge about the nature of certain mental events.

Note that to accept this much still leaves open what characteristics $E_1 \ldots E_n$ are. For all that has been said we should construe these as an experience's being the presentation of such and such mind-dependent qualities, as a sense-datum theory supposes. Or we might instead take them to be representational properties, as an intentional theory would press. For our present purposes here, we can remain neutral about this matter. All that matters for our current purposes is that such views will attribute to subjects who grasp the concept of perfect hallucination both the power to identify the marks of experience in having an experience and a recognition of their modal independence of the conditions of perceiving. To this extent, then, such theories are immodest in their attribution of epistemological powers that subjects have when they give an explanation of how we come to have a conception of sensory experience which can be employed from the first-person perspective.

This is not the only way to elaborate the initial sketch. Instead one may insist that the original instructions to conceive of perfectly matching hallucinations are all that is needed to give one a conception of perceptual experience. This second way of thinking about the idea of sense experience, we might call a modest or minimal conception. We need not look for some further characteristics in virtue of which an event counts as an experience of a street scene, but rather take something to be such an experience simply in virtue of its being indiscriminable from a perception of a street scene. Nothing more is needed for something to be an experience, according to this conception, than that it satisfy this epistemological condition. Rather than appealing to a substantive condition which an event must meet to be an experience, and in addition ascribing to us cognitive powers to recognise the presence of this substantive condition, it instead emphasises the limits of our powers of discrimination and the limits of self-awareness: some event is an experience of a street scene just in case it couldn't be told apart through introspection from a veridical perception of the street as the street.

Does this second conception really capture what we need? Well a proponent of the immodest view cannot fault a modest account for failing to capture in its conception of what a sense experience is all those situations that the immodest account deems to be perceptual experiences of a street scene. After all, by immodest lights the kind of experience one has when seeing such a street scene is of just the same kind as any non-perceptual event which is not a perception but still an experience as of

a street scene, namely an event with the properties $E_1 \ldots E_n$. Since nothing can be discriminated from itself, the immodest approach will hold that the modest one should agree that these events are indiscriminable from a veridical perception of a street scene and hence are perceptual experiences as of a street scene. (Of course, by modest lights this consequence might not follow, but that would only be because the particular version of the immodest account is inadequate and the properties $E_1 \ldots E_n$ it specifies are not after all sufficient for an event's being a perceptual experience as of a street scene even in the case of veridical perception.) So immodest views may complain that modest ones fail to capture what defines an event's being an experience but not that their conception of experience is too narrow.

On the other hand, it is difficult for an immodest account to avoid complaining that modest ones are far too catholic in their conception of what can be an experience as of a street scene. Given all we have said so far, nothing rules out as possible a situation in which $E_1 \ldots E_n$ are absent but in which a subject would be unable to discriminate through reflection this situation from one in which a street scene was really being seen. For the immodest view in question, this could not be a case of visual experience as of a street scene, while by modest lights that would be exactly what it is.

Now surely this result would be unfortunate for any immodest view, given our initial assumptions. For we supposed that reflection on experience offers support to a Naïve realist construal of sensory experience. When one reflects on one's experience it seems to one as if one is thereby presented with some experience-independent elements of the scene before one as constituents of one's experience and not merely as represented to one as in imagination. Even if the experience does also possess the characteristics $E_1 \ldots E_n$, it need not manifest to the perceiver that these are present as opposed to Naïve realist aspects of experience. And it is at least not manifest that the experience is the kind of experience it is in virtue of the presence of these properties as opposed to being Naïve realist—for were it, then clearly it would not even seem to us as if Naïve realism is true. When we turn to a case of perfect hallucination, we know that the Naïve phenomenal properties which seem to be present in the case of veridical perception certainly cannot be present in the case of hallucination. Of course they may still seem to be present, and in as much as the hallucination is indistinguishable from the perception they will seem to be so. So, if the presence of $E_1 \ldots E_n$ as opposed to the presence of Naïve phenomenal properties is not manifest to us in the case of veridical perception, and anyway is certainly not presented as definitive of that's being the experience it is, then it seems plausible that what links the case of hallucination to the veridical perception is the seeming presence of Naïve phenomenal properties and not $E_1 \ldots E_n$. In that case, common sense has no reason to discriminate against a case of perfect hallucination which lacks $E_1 \ldots E_n$ but yet which seems to possess the properties relevant to its being an experience as of a street scene in the first place, the seeming presence of Naïve phenomenal properties.

If a modest account is too catholic in its conception of experience, immodest views will seem from a common-sense perspective to be too restrictive. Even if the presence of $E_1 \ldots E_n$ is sufficient to determine that one is having an experience as of a street scene, nothing has shown why it has to be necessary. Rather, if it really is possible to produce an experience lacking those features but otherwise being indiscriminable from a perception of a street scene, the account will offer just one way in which such an experience can occur. A proponent of an immodest view can only hope to offer necessary as well as sufficient conditions for having an experience—and hence to explain the having of experience in terms of its favoured conditions—if it can ensure that the modest approach and its favoured form of immodesty coincide in the extension they give the concept of experience.

In turn, this coincidence of extension can be guaranteed only if the proponent of the immodest account embraces a substantive epistemic principle. That will be achieved only if the situation sketched above turns out to be impossible: that there cannot be any situation which is indistinguishable for its subject from actually perceiving a street scene and yet which lacks the relevant properties. In turn, one must assume that a subject couldn't but be in a position to discriminate a situation which lacked $E_1 \ldots E_n$ from one which possessed them. Here I just assume that for one situation to be indiscriminable from another requires only that it not be possible to know that it is distinct in kind.[11] Therefore to deny it is possible that a situation which is distinct in kind from an event possessing $E_1 \ldots E_n$ is not possibly knowable as distinct in kind, is to claim that for any situation distinct

in kind from an event possessing $E_1 \ldots E_n$ it is possible to know that it is distinct.

Adopting this position is to attribute a privileged epistemic position to the subject of experience. For, according to it, a responsible subject who wishes to determine how things are with him or herself through reflection must not only correctly identify phenomenal properties of a specific sort when they are present, but also they cannot be misled into judging them present when they are not. It is not merely that the properties which determine an event as an experience are held to be self-evident on this view—that the presence of such properties indicates to the subject that they are present when they are present. It must also be the case that the absence of such properties when they are absent is equally detectible by the subject, so that there is always some way that a subject could tell that he or she was not so experiencing when not doing so. It is to attribute to responsible subjects' potential infallibility about the course of their experiences.

Of course, some philosophers have assumed that these epistemic properties are definitional of the mental, and so see nothing substantive in the additional assumption. But the doctrine of infallibilism about the mental is particularly problematic in relation to sensory states once we are forced to admit that appearances systematically appear to us other than they are. For if we can be misled with respect to some properties of sensory experiences, there is a question as to what can motivate the claim that we are infallible in other judgements about them. As I indicated above, part of the motivation for disjunctivism is precisely the thought that introspection of our sense experience supports Naïve Realism, and hence forces us to see both sense-datum and intentional theories as forms of error theory.

The assessment of this epistemological commitment I'll leave for elsewhere. For this discussion, the only point to note is that given the need to rely on this assumption, an immodest approach to perceptual experience carries more theoretical burdens than does a modest approach. The burden of proof is not on the disjunctivist to show that we should adopt a less than conjunctive theory of appearances, the burden is really on any common kind theorist, to show that the theory they propose is not really too restrictive; or that the added epistemological burdens which come with demonstrating that are ones that we should accept.

This points to where Dancy was misled. In fixing on the concept of perceptual experience in general we seem to have no more resources than that we need to pick out something indiscriminable from veridical perception. So the most inclusive conception we can have here is an implicitly relational one. Any of the non-relational specifications that Naïve realist, or sense-datum, or intentional theories or some other approach can give us would seem just to offer at best a sufficient condition for meeting the relational specification. That would offer simply an account of one particular variant of experience, rather than an account of what experience must be. What most Common Kind theories (i.e. theories which endorse the Common Kind Assumption) ignore is that in giving an account of experience they normally succeed, if at all, only in giving sufficient conditions for one's experience to be a certain way, and fail to show that the conditions they offer are necessary. Without the latter being fulfilled, no such theory can claim to give a fully general account of experience.

Hence we can see that as long as our focus is on the concept of sensory experience in general, intended to cover all possible cases of what we would count as a sensory experience of a lavender bush, then our default position should be that of the disjunctivist. What we mean by this is no more than this is a situation which is indiscriminable through reflection from a veridical perception of a lavender bush.[12]

REFERENCES

Alston, W., "Back to the Theory of Appearing," *Philosophical Perspectives* (Epistemology) 13 (1999): pp. 181–203.

Anscombe, G. E. M., "The Intentionality of Sensation: A Grammatical Feature," in *Analytic Philosophy,* second series, R. Butler, ed., (Oxford: Blackwell, 1962).

_____ "Causality and Determination," in *Metaphysics and the Philosophy of Mind: Collected Papers,* vol. II (Oxford: Blackwell, 1981)/

Broad, C.D. *The Mind and Its Place in Nature* (London: Kegan Paul, 1925).

Chisholm, R. *Perception* (Ithaca, NY: Cornell University Press, 1959).

Dancy, J., "Arguments from Illusion," *The Philosophical Quarterly* 45 (1995): pp. 421–38.

Evans, G. *The Varieties of Reference*, J. McDowell. ed., (Oxford: Clarendon Press, 1982).

Firth, R., "Sense-Data and the Percept Theory," in *Perceiving, Sensing and Knowing*, R. Swartz, ed., (Los Angeles/Berkeley: University of California Press, 1965).

Foster, J. *A. J. Ayer* (London: Routledge, 1986).

_____ *The Nature of Perception* (New York: Oxford University Press, 2000).

Gendler, T. S. "Personal Identity and Thought Experiments," *The Philosophical Quarterly* 52, no. 206 (2002): pp. 34–54.

Graff, D., "Phenomenal Continuua and the Sorites," *Mind* 110, no. 440 (2001): pp. 905–35.

Harman, G., "The Intrinsic Quality of Experience," in *Philosophical Perspectives*, vol. 4, J. Tomberlin, ed., (Atascadero, CA: Ridgeview Publishing Co., 1990).

Hinton, J. M., "Visual Experiences," *Mind* 76 (1967): pp. 217–27.

_____ *Experiences: An Inquiry into Some Ambiguities* (Oxford: Clarendon Press, 1973).

Jackson, F. *Perception: A Representative Theory* (Cambridge: Cambridge University Press. 1977).

Langsam, H., "The Theory of Appearing Defended," *Philosophical Studies* 87 (1997): pp. 33–59.

Martin, M. G. F. 'Beyond Dispute,' in *The History of the Mind-Body Problem*, T. Crane and S. Patterson, eds., (London: Routledge, 2001).

_____ "The Transparency of Experience," *Mind and Language* 17, no. 4 (2002): 376–425.

McDowell, J., "Criteria, Defeasibility and Knowledge," *Proceedings of the British Academy* (1982).

_____ "Singular Thought and the Extent of Inner Space,' in *Subject, Thought and Context*, P. Pettit and J. McDowell, eds., (Oxford: Clarendon Press, 1986).

Merleau-Ponty, M. *La Structure de comportement*, A. Fisher, trans., (Paris: Presses Universitaires de France, 1942).

Moore, G. E., "The Nature of Perception???," in *Selected Papers*, T. Baldwin, ed., (London: Routledge, 1905). (Original edition, 1905).

_____ "The Refutation of Idealism," in *Philosophical Studies* (London: Routledge & Kegan Paul, 1922).

O'Shaughnessy, B. *The Will*, 2 vols. (Cambridge: Cambridge University Press, 1980).

Peacocke, C. A. B. *Sense and Content* (Oxford: Clarendon Press, 1983).

_____ *A Study of Concepts* (Cambridge, MA: MIT Press, 1992).

_____ "Externalist Explanation," *Proceedings of the Aristotelian Society* XCIII, (1993): pp. 203–30.

Price, H. H. *Perception* (London: Methuen, 1932).

Prichard, H. A. *Knowledge and Perception* (Oxford: Clarendon Press, 1950).

Robinson, H., "The General Form of the Argument for Berkeleian Idealism," in *Essays on Berkeley: A Tercentennial Celebration*, J. Foster and H. Robinson, eds., (Oxford: Clarendon Press, 1985).

_____ *Perception* (London: Routledge, 1994).

Russell, B. *The Problems of Philosophy* (Oxford: Oxford University Press, 1912), 9th ed.

Searle, J. *Intentionality* (Cambridge: Cambridge University Press, 1983).

Snowdon, P. F., "Perception, Vision and Causation," *Proceedings of the Aristotelian Society* (1980–81).

Sturgeon, S., "Visual Experience," *Proceedings of the Aristotelian Society* 98 (1998): pp. 179–200.

_____ *Matters of Mind: Consciousness, Reason and Nature* (New York: Routledge, 2000).

Tye, M. *Ten Problems about Consciousness* (Cambridge, MA: MIT Press, 1995).

Valberg, J. J. *The Puzzle of Experience* (Oxford: Clarendon Press, 1992).

Williamson, T. *Identity and Discrimination* (Oxford: Basil Blackwell, 1990).

_____ "Is Knowing a State of Mind?" *Mind* 104, no. 415 (1995): pp. 533–65.

_____ *Knowledge and Its Limits* (New York: Oxford University Press, 2000).

Yablo, S., "Mental Causation," *Philosophical Review* 101, no. 2 (1992): pp. 245–80.

_____ "Wide Causation," in *Philosophical Perspectives, 11, Mind, Causation, and World*, E. James, ed., (Boston: Blackwell, 1997).

NOTES

1. Recent defenders of intentionalism include Harman 1990; Peacocke 1983, 1992; Searle 1983, and Tye 1995. In the analytic tradition its popularity can be traced back to Firth's discussion of the percept theory in the mid-century (Firth 1965), on the one hand, and Anscombe's critique of both sense-datum theorists and their ordinary language opponents (Anscombe 1962), on the other. With some caveats, one can also see it as dominant within the phenomenological tradition.
2. To this purpose, the term 'sense-datum' was introduced first by Moore in Moore 1905 and made public in Russell 1912. Though out of favour in recent years, one can find defences of sense-data in Jackson 1977, O'Shaughnessy 1980, Foster 1986, and Robinson 1994. Subjectivism as here conceived captures a broader range of theories than just this, though, and includes for example the appeal to sensational properties in Peacocke 1983.
3. William Alston has recently defended a theory of appearing while claiming of hallucinations that we can consider them to be awarenesses of mental images, see Alston 1999, pp. 191–92. He suggests that nothing positively shows that mental images are dependent on our awareness of them, and if one could maintain this conclusion, the argument of the text would be blocked. However he does not discuss what model of the causation of hallucination we would then need to adopt: can the local conditions for producing mental images be sensitive to the absence of an external object of perception? If not, which is the overwhelming plausible conclusion to

draw, then if the veridical perception is the same kind of mental state, we will get the conclusion drawn in the text.

4. Harold Langsam, who endorses disjunctivism, seeks to block the argument from hallucination by suggesting that there are possible accounts of hallucination on which a hallucination is, for example, a relation to the region of physical space where an object appears to be (Langsam 1997, p. 47). However, Langsam's agnostic stance about the nature of hallucinations is misleading about the force of the argument against the Naïve Realist. Of course there may be some hallucinations which are examples of awareness of the mere air around us. But the pressing question is whether there are any which take the form indicated in the text and which are of the same kind as veridical perceptions. Langsam does nothing to show that such experiences are impossible, nor does he discuss the consequences of the possibility of their existence.
5. Such talk of intentional objects can be traced at least to Anscombe 1962, who claims Medieval authority for it, and this way of expressing the view is echoed in Harman 1990. Nonetheless, critics tend to read the talk as involving a commitment to a special kind of entity, which unsurprisingly leads to a dismissal of mystery mongering. No such ontological profligacy need be, or was intended by those who chose to talk in this way.
6. As I propose at greater length in Martin 2001.
7. See in particular Martin 2002.
8. This is true of Moore and Russell, who insisted that the objects of sensing must be independent of our awareness of them (see Moore 1922, Russell 1912, Broad 1925, Price 1932). For an early criticism of precisely this aspect of the sense-datum tradition (Prichard 1950).
9. Merleau-Ponty 1942; Valberg 1992.
10. However, Hannah Ginsborg reminded me that Hinton does discuss fact perception in the later monograph, *Experiences*, see pp. 101–124; so the contrast is perhaps not as stark as I present it here.
11. Here I follow the approach to indiscriminability found in Williamson 1990; see below for further discussion of the relevant properties of indiscriminability.
12. This paper originated as a twenty-minute talk at CREA in Paris and a written draft was produced during a visit at the RSSS of the ANU; versions of the paper have been read to audiences in Paris, Canberra, Dubrovnik, Edinburgh, London, Leeds, Helsinki, and Oberlin. I am grateful in particular for comments on this material to Tim Crane, Alan Hajek, Jen Hornsby, Véronique Munoz-Dardé, Panu Raatikainen, Susanna Siegel, Paul Snowdon, Maja Spener, Charles Travis, and above all to Scott Sturgeon for provoking much of the second half of the paper.

Is the Visual World a Grand Illusion?

Alva Noë

Traditional scepticism about perceptual experience questions whether we can know that things are as we experience them as being. This paper targets a new form of scepticism about experience that takes its start from recent work in perceptual psychology and philosophy of mind. The new scepticism questions whether we even have the perceptual experience we think we have. According to the new scepticism, we have radically false beliefs about what our perceptual experience is like. Perceptual consciousness is a kind of false consciousness; a sort of confabulation. The visual world is a grand illusion.

The new scepticism raises important questions for philosophy, psychology, and consciousness studies. What is the character of our perceptual experience? And who does the sceptic mean by 'we' anyway? Ordinary perceivers? Ordinary perceivers in unusual reflective contexts? Or psychologists and philosophers? These are surprisingly difficult questions. I argue, in what follows, that the new scepticism, and perhaps also the new perceptual psychology it has spawned, rests on a misguided and overly simplistic account of perceptual phenomenology.

I.

According to a conception of visual experience that has been widely held by perceptual theorists, you open your eyes and—*presto!*—you enjoy a richly detailed picture-like experience of the world, one that represents the world in

From *Journal of Consciousness Studies* 9, no. 5–6 (2002): pp. 1–12.

sharp focus, uniform detail and high resolution from the centre out to the periphery. Let us call this the snapshot conception of experience.

Empirical investigation of the nature of vision takes its start from the snapshot conception. The puzzle visual theory faces is that of understanding how it is we come to enjoy such richly detailed snapshot-like visual experiences when our actual direct contact with the world in the form of information on the retina is so limited. The limitations are familiar: there are two retinal images, not one, and they are distorted, tiny, and upside-down (Gregory 1966–97, p. 1). In addition, the resolving power of the eye is limited and nonuniform; outside the high-resolution foveal region, the retina is nearly colour-blind and its powers of discrimination are severely limited. On top of this, the eye is in nearly constant motion, saccading from point to point in the visual field three or four times a second. As a result of saccadic suppression, the data made available to the retina takes the form of a succession of alternating snapshots and grey-outs.

How, on the basis of this fragmented and discontinuous information, are we able to enjoy the impression of seamless consciousness of an environment that is detailed, continuous, complex and high-resolution? This is *the* problem faced by visual theory.

The orthodox strategy is to suppose that the brain integrates information available in successive fixations into a stable, detailed model or representation. This stable representation then serves as the substrate of the actual experience. According to this orthodox approach, vision just is the process whereby the patchy and fragmentary bits of information on the retina are transformed into the detailed stable representations underlying actual perceptual experience. This is what David Marr had in mind, I think, when he wrote that 'Vision is the process of discovering from images what is present in the world, and where it is' (Marr 1982, p. 3).

II.

Recent work in perceptual psychology challenges this traditional framing of the problem for visual theory by questioning whether we really enjoy the sort of richly detailed, snapshot-like visual experiences we think we do. If we do not enjoy such experiences, then we are not faced with the problem of how the brain gives rise to them. Indeed, from the standpoint of what I am calling the new scepticism, the central problem of visual theory is not: how do we see so much on the basis of so little? It is, rather, why does it seem to us as if we see so much when in fact we see so little?

The point is beautifully epitomized by Dennett, who is the *éminence grise*, and strongest proponent, of the new scepticism. Edelman had written 'One of the most striking things about consciousness is its continuity' (1989, p. 119). Dennett writes in response:

> This is utterly wrong. One of the most striking features about consciousness is its discontinuity—as revealed in the blind spot, and saccadic gaps, to take the simplest examples. The discontinuity of consciousness is striking because of the *apparent* continuity of consciousness (1991, p. 356).

This remark is wonderful because it makes very clear that the worry is about the nature of experience or consciousness itself. We are misled as to the true nature of consciousness, Dennett is saying. Consciousness is really discontinuous. It appears to us to be continuous. A paradoxical way to put the point would be: it turns out that we are mistaken in our assessment of how things seem to us be.

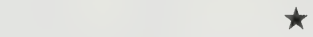
III.

How does the argument for the new scepticism about experience go? What is the argument that experiences are not what they seem to be? The *locus classicus* is Dennett's discussion of filling in at the blind spot (Dennett 1991, pp. 344–56).[1]

There is a blind spot in each eye in the sense that there is a place on each retina where there are no photoreceptors. We don't usually notice the blind spot. What falls on the blind spot of one eye doesn't fall on the blind spot of the other, and the eyes are in nearly constant motion anyway, so what falls on the blind spot now doesn't fall on the blind spot a moment later. But you don't experience a hole in your visual field even when you stare with one eye at a white wall (say). It takes special care to demonstrate the existence of the blind spot. Shut your right eye and fixate the star below. If you move the page to the right distance from your face (about 8–12 inches), you will be unable to see the black disc on the left. The black disc disappears because it falls in your blind spot.

Demonstrations like this are frequently cited as evidence that the brain *fills in* the gap in our internal representation of the visual field (e.g.,

Palmer 1999, p. 617). How else can you explain the phenomenon? Dennett noticed that the fact that we do not experience a gap in the visual field corresponding to the blind spot does not entail that the brain fills in the gap. This discounts other possibilities, such as that the brain simply ignores the blind spot. If the brain ig-



nores the absence of information from the part of the field corresponding to the blind spot, then it doesn't represent that information as absent. But then there is nothing to be filled in. Even if the brain *does* represent the absence of information, it isn't obvious that it must fill the missing information in. After all, if the brain knows what it needs to fill in, then for whose benefit is the operation of filling in performed? The brain's job is finding out, Dennett asserts, not filling in. In the absence of direct evidence of the process of filling in itself, and not merely of the putative effects of filling in—namely, a gap-free experience—we aren't entitled to suppose that filling in occurs.

What does this have to do with the new scepticism? Dennett seems to have believed that there is no such good evidence of processes of filling in.[2] Let's grant him this assumption. The interesting bit is what he takes to follow from this. If there is no filling in at the blind spot, then, he reasons, there must be a gap in our experience of the visual world; a gap which, however, we fail to notice. This, presumably, is an example of the discontinuity of experience despite its apparent continuity. We take our experience to be gap-free when it is not. We are the victims of an illusion of visual consciousness.

IV.

But does this sceptical reasoning go through? It is certainly right that you don't notice a gap in the visual field corresponding to the blind spot even under monocular viewing conditions. In general, if you shut your eye and stare at the wall, you have a visual experience as of a gap-free expanse of the wall. That is, it looks to you as if there is an unbroken expanse of wall. But this is not to say that it seems to you as if, as it were in a single fixation, you experience *the whole of the wall's surface*. If you reflect on what it is like for you to look at the wall, you will notice that it seems to you as if the whole wall is there at once, but not as if every part of the wall's surface is represented in your consciousness at once. Rather, you experience the wall as present and you experience yourself as having access to the wall by looking here, or there, attending here, or there. It is no part of ordinary phenomenology that we experience the whole wall, every bit of it, in consciousness all at once.[3]

The sceptical argument seems to turn on attributing to us, as lay perceivers, something like the snapshot conception of experience. According to this conception, visual experiences are like snapshots that represent the scene in high-resolution focus and sharp detail. Dennett then points out, convincingly, that our experience is not like a snapshot—there's a blind spot, bad parafoveal vision, etc.—and he concludes that we are victims of an illusion about the character of our own consciousness.

But the mistake in question—the snapshot conception of experience—is not one to which perceivers themselves are committed. Perhaps it is an idea about perception that psychologists or philosophers find natural. Perhaps it is way of describing experience that many ordinary perceivers would be inclined to assent to if they were asked appropriately leading questions. But this is compatible with its being the case that we do not really take our experience to be this way.

V.

A second important source for the new scepticism is recent work on change blindness and inattentional blindness in the psychology of scene perception.[4]

To set the stage, consider the following familiar sort of gag. I say to you as you tuck into your lunch: 'Hey? Isn't that Mick Jagger over there?' You turn around to look. When you do, I snatch one of your French fries. When you turn back, you're none the wiser. You don't remember the exact number or layout of fries on your plate and you weren't paying attention when the fry was snatched. Your attention was directed elsewhere.

It turns out—this is the central finding of work on change blindness conducted by O'Regan, Rensink, Simons, Levin, and others[5]—that this sort of failure to notice change is a pervasive feature of our visual lives. Usually, when changes occur before us, we notice them because our attention is grabbed by the flickers of movement associated with the change. But if we are prevented from noticing the flicker of movement when the change occurs, say because at the same

time flickers occur elsewhere, we may fail to notice the change (O'Regan *et al.*, 1996, 1999). What is striking—and this will become important later on—is the fact that we will frequently fail to notice changes even when the changes are fully open to view. Even when we are looking right at the change when it occurs, something we can test with eye trackers, we may fail to see the change (O'Regan *et al.*, 2000).

The fact of change blindness is widely thought to have several important consequences. First, perception is, in an important sense, attention-dependent. You only see that to which you attend. If something occurs outside the scope of attention, even if it's perfectly visible, you won't see it. In one study, perceivers are asked to watch a video tape of a basketball game and they are asked to count the number of times one team takes possession of the ball (Neisser 1976, Simons and Chabris 1999). During the film clip, which lasts a few minutes, a person in a gorilla suit strolls onto the centre of the court, turns and faces the audience and does a little jig. The gorilla then slowly walks off the court. The remarkable fact is that perceivers (including this author) *do not* notice the gorilla. This is an example of inattentional blindness.[6] Second, perception is gist-dependent. Some changes, for example, in the features that affect the gist of the scene, are more likely to be noticed (Simons and Levin 1997). Third, it seems that the brain does not build up detailed internal models of the scene; that is, it doesn't perform the integration of information across successive fixations, contrary to the assumption of traditional orthodoxy (Blackmore *et al.* 1995, Rensink *et al.* 1997, O'Regan *et al.* 1999, Rensink *et al.* 2000, Noë *et al.* 2000). Or if it does, we have little easy access to this detail. If we did, then presumably we'd keep track of change better than we do.

VI.

Many of the investigators on change blindness believe that this work supports the grand illusion hypothesis. For example, Susan Blackmore and her colleagues 1995, p. 1075, write:

> we believe that we see a complete, dynamic picture of a stable, uniformly detailed, and colourful world, but [o]ur stable visual world may be constructed out of a brief retinal image and a very sketchy, higher-level representation along with a pop-out mechanism to redirect attention. The richness of our visual world is, to this extent, an illusion.

In a similar vein, O'Regan 1992, p. 484, writes:

> despite the poor quality of the visual apparatus, we have the subjective impression of great richness and 'presence' of the visual world. But this richness and presence are actually an illusion. . .[7]

The problem with this reasoning is the same as we saw above in connection with Dennett's discussion of the blind spot. It just is not the case that we, normal perceivers, believe we see a complete, dynamic picture of a stable, uniformly detailed and colourful world. Of course it *does* seem to us as if we have perceptual access to a world that is richly detailed, complete and gap-free. And we do! We take ourselves to be confronted with and embedded in a high-resolution environment. We take ourselves to have access to that detail, not all at once, but thanks to movements of our eyes and head and shifts of attention.[8]

Consider a question posed by Rensink *et al.* 2000, p. 28: 'Why do we feel that somewhere in our brain is a complete, coherent representation of the entire scene?' But this question rests on a false presupposition. It does not seem to us as if somewhere in our brain there is a complete, coherent representation of the scene. Perceptual experience is directed to the world, not to the brain.

VII.

If I am right that perceivers are not committed to the idea that they have detailed pictures in the head when they see (the snapshot conception), then how can we explain the fact that perceivers are surprised by the results of change blindness? Does not the surprise itself register our commitment to the problematic, snapshot conception of experience? This objection has been raised by Dennett 2001 (see also Dennett):

> why do normal perceivers express such surprise when their attention is drawn to [the relevant facts about their perceptual limitations]. Surprise is a wonderful dependent variable, and should be used more often in experiments; it is easy to measure and is a telling betrayal of the subject's *having expected something else*. These expectations are, indeed, an overshooting of the proper expectations of a normally embedded perceiver–agent; people shouldn't have these expectations, but they do. People are shocked, incredulous, dismayed; they often laugh and shriek when I demonstrate the effects to them for the first time. These behavioral responses are themselves data in good standing, and in need of an explanation.

This is an important objection, but one that is easy to answer. The astonishment people experience when confronted with the facts of change blindness and inattentional blindness does indeed demonstrate that their beliefs are upset by these demonstrations. But one need not attribute to them (to us) a commitment to the snapshot conception. The surprise is explained simply by supposing that we tend to think we are better at noticing changes than in fact we are, or that we are much less vulnerable to the effects of distracted attention than we in fact are. This is a plausible explanation of the surprise we feel when confronted with the results, and one that does not foist on us the ideology of the snapshot conception.

Surprise requires explanation, but so does the lack of surprise. Notice that we are not surprised or in any way taken aback by our need to move eyes and head to get better glimpses of what is around us. We peer, squint, lean forward, adjust lighting, put on glasses, and we do so automatically. The fact that we are not surprised by our lack of immediate possession of detailed information about the environment shows that we don't take ourselves to have all that information in consciousness all at once. If we were committed to the snapshot conception, wouldn't we be surprised by the need continuously to redirect our attention to the environment to inform ourselves about what is there?

Finally, it is worth noting that artists, magicians, stage designers and cinematographers—people who live by the maxim that the hand is quicker than the eye—would not be surprised by the change blindness results. Why should they be? Our perceptual access to the world is robust, but fallible and vulnerable. How could one really think otherwise?[9]

VIII.

Let us summarize what we have found so far. First, the new scepticism is right about some things. For example, it is right that experience does not conform to the snapshot conception. And so it is right that visual science should not concern itself with how the brain produces experiences thought of like that. But the new scepticism seems to rest on a substantially false characterization of what perceptual experience actually seems to us—that is, to lay perceivers—to be like. In particular, it attributes to us something like the snapshot conception. The scepticism can be resisted if we recognize that we are not committed to the snapshot conception. We don't take ourselves to experience all environmental detail in consciousness all at once. Rather, we take ourselves to be situated in an environment to have access to environmental detail as needed by turns of the eyes and head, and repositioning of the body.

IX.

But we are not done yet. We must not be too quick in dismissing the grand illusion hypothesis. One of the results of change blindness is that we only see, we only experience, that to which we attend. But surely it is a basic fact of our phenomenology that we enjoy a perceptual awareness of at least some unattended features of the scene. So, for example, I may look at you, attending only to you. But I also have a sense of the presence of the wall behind you, of its colour, of its distance from you. It certainly seems this way. If we are not to fall back into the grip of the sceptic's worry, we must explain how it is we can enjoy perceptual experience of unattended features of a scene. Let us call this the problem of perceptual presence.

The problem of perceptual presence forces us to confront the grand illusion puzzle again. But this version of the sceptical worry is stronger, for it does not rely on the misattribution to us of the phenomenologically inadequate snapshot conception of experience. All that it requires is that we acknowledge that we are perceptually aware, sometimes, of unattended detail. And who could deny that?

We can sharpen the worry. One of the main upshots of work on change blindness is that the brain does not produce a detailed world model corresponding to perceived detail. The sceptical problem then becomes: how can we enjoy experiences of the world as richly detailed when we lack internal representations of all that detail?

X.

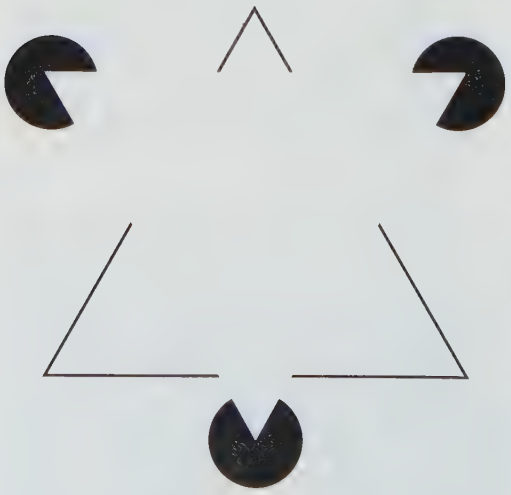
To begin to see our way clear to a solution of the problem of perceptual presence, consider as an example a perceptual experience such as that you might enjoy if you were to hold a bottle in your hands with eyes closed. You have a sense of the presence of a whole bottle, even though you only make contact with the bottle at a few isolated points. Can we explain how your experience in this way outstrips what is actually

given, or must we concede that your sense of the bottle as a whole is a kind of confabulation?

Or consider a different case: there is a cat sitting motionless on the far side of a picket fence. You have a sense of the presence of a cat even though, strictly speaking, you only see those parts of the cat that show through the fence. How is it that we can in this way enjoy a perceptual experience as of the whole cat?

One way we might try to explain this is by observing that you draw on your knowledge of what bottles are, or what cats are. You bring to bear your conceptual skills. This is doubtless right. But it does not, I think, do justice to the phenomenology of the experience. For crucially, your sense of the presence of the bottle is a sense of its *perceptual* presence. That is, you do not merely *think* or *infer* that there is a bottle present, in the way, say, that you think or infer that there is a room next door. The presence of the bottle is not inferred or surmised. It is *experienced*. And so with the cat: you see it there, you experience it, even though you only see parts of it.[10]

This is an example of what psychologists call *amodal perception*. As an illustration, consider the famous Kanisza figure (left). Most perceivers take themselves to experience two triangles, one of which is above, and so partly blocks from view, the other. In addition, the topmost triangle partially covers the three black disks. The hidden portions of the disks and the lower triangle are said to be *amodally* perceived as complete. Here you experience as perceptually present something which is, in fact, hidden from view.



Amodal perception is an important phenomenon. It is involved in our perception of solidity, as, for example, when you experience a tomato as three-dimensional and round, even though you only see its facing side, or when you experience a chair as whole and intact, even though it is partially blocked from view by the table.

Amodal perception is paradoxical in that it is perceiving what is, strictly speaking, out of view. I would like to suggest that we approach the problem of perceptual presence as, in essence, a problem about amodal perception. The proposal—this is a step toward the solution of the problem of perceptual presence—is that the detail of the world is present to consciousness, but in the way that amodally perceived features of scenes or objects are amodally present. They are perceived without being really perceived. The question whether the visual world is a grand illusion then transposes itself into the question whether amodal perception should be thought of as illusory.

XI.

Traditional orthodoxy speaks to the problem of perceptual presence by supposing that we build up an internal model corresponding to the experienced detail. But this sort of approach faces obstacles. As we have noticed, work on change blindness seems to suggest that we may not in fact actually produce such detailed internal models.

But consider a more basic point: why should the brain go to the trouble of producing a model of the bottle when the bottle is right there in your hands and can serve as a repository of information about itself? All the information about the bottle you need is available to you in the world—you need only move your hands to gather it. And so for the cat. Why represent the cat in all its detail when all the information you need is available, when you need it, by eye and head movements?[11]

I think that what makes the orthodox move seem so attractive is that theorists tend to rely on a snapshot conception of experience according to which we take ourselves in experience to represent the cat or the bottle in consciousness in all its detail. But this distorts the phenomenology. It does not seem to me as if every part of the cat is visible to me now, even though it

does seem to me now as if I perceive a whole cat and as if the unperceived parts of the cat's body are present. After all, I can *see* that the cat is partly hidden behind the fence! This is just the thing with amodal perception: one experiences the presence of that which one perceives to be out of view.

XII.

The solution to the problem of perceptual presence is achieved in two steps.[12] First, we need to reflect more carefully on the phenomenology. When we do so, it becomes clear that our sense of the presence of the cat as a whole now does not consist in our representation, now, of the whole of the cat in consciousness. It consists rather in the fact that we *now* have access to the whole of the cat. Second, the basis of this access is our possession of sensorimotor skills (O'Regan and Noë 2001a). In particular, its basis is those skills—practical knowledge of the ways what we do gives rise to sensory stimulation—whose possession is constitutive of sensory perception. My relation to the cat behind the fence is mediated by such facts as, when I blink, I lose sight of it altogether, but when I move a few inches to the right, a part of its shoulder that was previously hidden comes into view. My sense of the perceptual presence, now, of that which is now hidden behind a slat in the fence consists in my expectation that by movements of the body I can produce the right sort of new cat stimulation.

In general, our sense of the perceptual presence of the detailed world does not consist in our representation of all the detail in consciousness now. Rather, it consists in our access now to all of the detail, and to our knowledge that we have this access. This knowledge takes the form of our comfortable mastery of the rules of sensorimotor dependence that mediate our relation to our immediate environment. My sense of the presence of the whole cat behind the fence consists precisely in my knowledge, my implicit understanding, that by a movement of the eye or the head or the body I can bring bits of the cat into view that are now hidden. This is one of the central claims of the enactive or sensorimotor approach to perception (O'Regan and Noë 2001a, Noë forthcoming).[13]

XIII.

Note: my sense of the presence of the hallway next door is not in this way mediated by patterns of sensorimotor dependence (O'Regan and Noë 2001a). I can jump up and down, turn around, turn the lights on and off, blink, and so on, and it makes no difference whatsoever to my sense of the presence of the room next door. My relationship to the room next door—however strongly I believe or know or assume that it is present—is not a perceptual relation. My relation to the cat, however, or to the bottle, is. It is my implicit understanding of this that gives me the feeling and that justifies me in the feeling that the cat and the bottle are present to me.[14]

XIV.

The enactive approach to perception—with its emphasis on the centrality of our possession of sensorimotor skills—provides the basis, then, for a satisfying reply to the sceptic, but only provided that we adopt a more plausible phenomenology of perceptual experience. On this more plausible account, it is not the case that we take ourselves to represent the whole scene in consciousness all at once. The enactive, sensorimotor account explains how it can be that we enjoy an experience of worldly detail which is not represented in our brains. The detail is present—the perceptual world is present—in the sense that we have a special kind of access to the detail, an access controlled by patterns of sensorimotor dependence with which we are familiar. The visual world is not a grand illusion.

Acknowledgements

The ideas in this paper grow out of my collaborations with Evan Thompson and Kevin O'Regan. I wish to make explicit my debt to them. Thanks also to audiences at UC Riverside, UC Irvine, Cal Tech, Brooklyn College, and Cal Arts, where I have presented this material. Thanks to Jeff Barrett, Sue Blackmore, Dave Chalmers, Tori McGeer, Dominic Murphy, Philip Pettit, Kyle Sanford, and Eric Schwitzgebel for helpful conversation. Finally, I would like gratefully to acknowledge the support of a University of California President's Fellowship in the Humanities and faculty research funds of the University of California, Santa Cruz.

REFERENCES

Blackmore, S. J., G. Brelstaff, K. Nelson, and T. Troscianko, "Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes," *Perception* 24 (1995): pp. 1075–81.

Brooks, R. A., "Intelligence without reason," in *Proceedings of the 1991 International Joint Conference on Artificial Intelligence* (1991), pp. 569–95.

Dennett, D. C. *Consciousness Explained* (Boston: Little, Brown & Co., 1991).

Edelman, G. *The Remembered Present: A Biological Theory of Consciousness* (New York: Basic Books, 1989).

Gregory, R. L. *The Intelligent Eye: The Psychology of Seeing* (Princeton, NJ: Princeton University Press, 1966/1997), 5th ed.

Mack, A., and Rock, I. *Inattentional Blindness* (Cambridge, MA: MIT Press, 1998).

Marr, D. *Vision* (New York: W. H. Freeman, 1982).

Neisser, U. *Cognition and Reality: Principles and Implications of Cognitive Psychology* (San Francisco: W. H. Freeman, 1976).

Noë, A., "Experience and the active mind," *Synthese* 129, no. 1 (2001): pp. 41–60.

_____ *Action in Perception* (Cambridge, MA: MIT Press, forthcoming.

Noë, A., and O'Regan, J. K., "Perception, attention and the grand illusion," *Psyche* 6, no. 15 (2000), http://psyche.cs.monash.edu.au/v6/psyche-6-15-noe.html.

Noë, A., L. Pessoa, and E. Thompson, "Beyond the grand illusion: what change blindness really teaches us about vision," *Visual Cognition* 7, no. 1/2/3 (2000): pp. 93–106.

O'Regan, J. K. "Solving the 'real' mysteries of visual perception: the world as an outside memory," *Canadian Journal of Psychology* **46** (1992): pp. 461–88.

_____ "Change blindness," in *Encyclopedia of Cognitive Science* (London: Macmillan, Nature Publishing Group, forthcoming).

O'Regan, J. K., H. Deubel, J. J. Clark, and R. A. Rensink, "Picture changes during blinks: looking without seeing and seeing without looking," *Visual Cognition* 7 (2000): pp. 191–212.

O'Regan, J. K., and Noë, A., "A sensorimotor account of vision and visual consciousness," *Behavioral and Brain Sciences* 24, no. 5 (2001a).

_____ "What it is like to see: A sensorimotor theory of perceptual experience," *Synthese* 129, no. 1 (2001b): pp. 79–103.

O'Regan, J. K., J. A. Rensink, and J. J. Clark, "'Mud splashes' render picture changes invisible," *Investigative Ophthalmology & Visual Science* 37 (1996): p. S213.

_____ "Change-blindness as a result of 'mudsplashes,'" *Nature* 398 (1999): p. 34.

Palmer, S. E. *Vision Science: Photons to Phenomenology* (Cambridge, MA: MIT Press, 1999).

Pessoa, L., E. Thompson, and A. Noë, "Finding out about filling in: a guide to perceptual completion for visual science and the philosophy of perception," *Behavioral and Brain Sciences* 21, no. 6 (1998): pp. 723–802.

Rensink, R. A., J. K. O'Regan, and J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science* 8, no. 5 (1997): pp. 368–73.

_____ "On the failure to detect changes in scenes across brief interruptions," in *Visual Cognition*, 7(1–3), 127–145.

Shimojo, S., Y. Kamitani, and S. Nishida, "Afterimage of perceptually filled-in surface," *Science* 293 (2001): p. 1677.

Simons, D. J., "Current approaches to change blindness," *Visual Cognition* 7, no. 1/2/3 (2000): pp. 1–15.

Simons, D. J., and Chabris, C. F., "Gorillas in our midst: sustained inattentional blindness for dynamic events," *Perception* 28, no. 9 (1999): pp. 1059–74.

Simons, D. J., and Levin, D. T., "Change blindness," *Trends in Cognitive Sciences* 1, no. 7 (1997): pp. 261–7.

_____ "Failure to detect changes to people in a real-world interaction," *Psychonomic Bulletin and Review* 5 (1998): pp. 644–9.

Thompson, E., A. Noë, and L. Pessoa, "Perceptual completion: a case study in phenomenology and cognitive science," in *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, ed. J. Pettitot, F. J. Varela, Pachoud, and J-M. Roy (Stanford, CA: Stanford University Press, 1999).

Varela, F. J., E. Thompson, and E. Rosch, *The Embodied Mind* (Cambridge, MA: MIT Press, 1991).

NOTES

1. See Pessoa *et al.* 1998 and Thompson *et al.* 1999 for a more detailed critical assessment of Dennett's account of filling in.
2. Pessoa *et al.* 1998 argued that, Dennett's claims to the contrary notwithstanding, there is in fact evidence of the process of filling in itself. However, we also argued that, once Dennett's critical observations are taken on board, filling in loses much of its theoretical importance. I won't revisit these issues here. It's worth mentioning, however, that recent work in the lab of Shinsuke Shimojo at Cal Tech seems to provide strong evidence of filling in. In particular, Shimojo and his colleagues show that amodally filled-in figures generate afterimages (Shimojo *et al.* 2001).
3. See Thompson *et al.* 1999 and Noë forthcoming for further development of this line of criticism.
4. For recent reviews of the change blindness literature, see O'Regan forthcoming, Simons 2000,

Simons and Levin 1997. For a discussion of philosophical implications, see Noë et al. 2000, Noë and O'Regan 2000, and O'Regan and Noë 2001a. See Mack and Rock 1998 for a thorough treatment of inattentional blindness.

5. O'Regan *et al.* 1996, 1999; Rensink *et al.* 1997, 2000; Simons and Levin 1998.
6. The term is due to Mack and Rock 1998. A detailed study of the phenomenon is contained in their book. For further discussion in connection with the themes of this paper, see Noë and O'Regan 2000.
7. O'Regan no longer defends the grand illusion hypothesis. See, for example, O'Regan and Noë 2001a.
8. For more on this line of criticism, see Noë *et al.* 2000, Noë and O'Regan 2000, and O'Regan and Noë 2001a.
9. An artist friend of mine, working on a portrait series, asked me to sit for him. I was struck by the frenzy of his looking-activity. The rendering proceeded by means of an uninterrupted pattern of looking back and forth from me to the canvas and then back again. The detail wasn't in his memory, or in his internal representations. It was to be found in his subject (in me).
10. See Thompson *et al.* 1999 for more on this distinction.
11. O'Regan 1992 makes this point. There is no need to represent the detail of the environment in memory because we can let the world serve as its own 'outside memory.' Brooks 1991 makes a very similar point: the world, he proposes, can serve as its own best model.
12. The solution to the problem of perceptual presence is developed in a series of papers I have written with Kevin O'Regan: O'Regan and Noë 2001a, b; Noë and O'Regan 2000. See also Noë 2001; forthcoming.
13. I borrow the term 'enactive' from Varela *et al.* 1991.
14. Of course, there are sensorimotor dependencies mediating my relation to the room next door as well. Indeed, no sharp line can be drawn between that which is amodally perceived as present and that which is merely thought of as inferred. This is a strength of the view I am defending here. It suggests a way in which thought is grounded in the sort of sensorimotor knowledge that is, on just about any view, shared by humans and other animals.

Which Properties Are Represented in Perception?

Susanna Siegel

In discussions of perception and its relation to knowledge, it is common to distinguish what one comes to believe on the basis of perception from the distinctively perceptual basis of one's belief. The distinction can be drawn in terms of propositional contents: there are the contents that a perceiver comes to believe on the basis of her perception, on the one hand; and there are the contents properly attributed to perception itself, on the other. Consider the content:

(#) that Ms. Elfenbein went to Ankara.

Suppose that you believed that Ms. Elfenbein is out of town only if she went to Ankara. And suppose you went to her house, and found that her curtains were drawn, that her mailbox was overstuffed, and that repeated ringings of her doorbell went unanswered. Then you would reasonably come to believe (#) on the basis of your visual experience. But (#) does not seem to be properly attributable to your visual experience itself (in this case). We seem to be able to distinguish what is presented perceptually from what we go on to believe. We can draw this distinction, no matter how much or how little overlap there may be between them.

The same point can be put in terms of the properties that are represented in visual experience. Consider the property of being round. This is a property that something can be seen to have. And if one can perceive that a surface is round, roundness, the property, can be represented in visual experience. Contents and properties, then, are related in the following straightforward way: if a subject S's visual experience has the content that a thing x is F, then S's visual experience represents the property of being F.

It is relatively uncontroversial that color and shape properties of some sort are represented in visual experience.[1] Being orange and being spherical, for example, are properties that we can sensorily perceive a basketball to have. (I'm assuming that if shape properties are

From Tamar S. Gendler & John Hawthorne, eds., *Perceptual Experience* (Oxford University Press, 2006).

represented in experience, then so are some depth properties.) I'll also be assuming, somewhat more controversially, that in experience we also represent ordinary objects. The class of ordinary objects is notoriously difficult to define, but it is clear enough to support theorizing by psychologists—for instance, theorizing about what *concept* of object infants have. And plenty of examples of ordinary objects can be given—cats, keys, tables, and the like.

The main question addressed in this chapter is whether any properties are represented in visual experience, besides the ones standardly taken to be so represented. Do any sensory experiences represent any properties other than color, shape, illumination, motion, their co-instantiation in objects and successions thereof? I will focus on visual experiences, and argue that some visual experiences do represent properties other than these. Although the properties other than these do not form a natural class, it will be useful to have a label for them. Because they include, though are not limited to, natural kind properties, and because one of my examples will involve such a property, and finally because 'kind' begins with 'k,' I'm going to call the rest of the properties *K-properties*.[2] The thesis I will defend is:

> Thesis K: In some visual experiences, some K-properties are represented.

Defining K-properties in the way I have brings into focus some alternatives to Thesis K. Consider the following extreme view about the properties represented in visual experience: visual experience is akin to what David Marr called the *2 1/2-D sketch*. Roughly, the 2 1/2-D sketch represents color, shape, and illumination properties of facing surfaces, but does not represent which surfaces belong to the same object, or how those surfaces continue out of view. (That's what the '1/2' is for: some facing surfaces are represented as farther away than others; but the sketch does not represent full volumetric information.)

A slightly more permissive view is that visual experience represents that some surfaces and edges—for example, those making up a cup's handle and the rest of the cup—are grouped together into fully volumetric (3D) units. More permissive still is the view that visual experience represents colors, shapes, volumetric groupings, and objects. Thesis K is even more extreme: it allows that in addition to all these things, visual experience represents properties such as being a house, and being a tree.[3]

The views I've just mentioned differ on what the veridicality conditions of visual experience are. The less committal the contents of visual experience, the less misperception there is. For instance, suppose you and your brother come across a bowl full of expertly designed wax fruits. Your brother is fooled into thinking that there are ripe juicy peaches and pears in the bowl: he *believes* that there are peaches and pears in the bowl, and this belief of his is false. The scene doesn't fool you, let's suppose, but only because you already believed on some non-perceptual basis—for instance, from reading your daily horoscope's predictions—that you would see some fake fruits today. Because you have this background belief, you suspect trickery, and, unlike your brother, you don't end up *believing* that there are peaches and pears in the bowl. Might there be in such a case some sort of error in your visual experience, even if not in your belief? A *perceptual* error would be one from which not even your suspicion protects you. If you misperceive, then your visual experience's content is false: your visual experience tells you that there are peaches and pears on the table, and that is incorrect, so the experience is falsidical. In contrast, if no perceptual error is involved in this case, then the contents of your visual experience are less committal, but correct: they tell you, for instance, that the contents of the bowl have certain colors and shapes. So if you misperceived, then, in suspecting that things were not as they looked, you corrected for an error at the level of visual experience. Whereas if your visual experience told you something less committal about what you each saw in the bowl (as it would, for instance, if Thesis K were false), then your suspicion saved you from making an error at the level of belief in the first place. These two descriptions of the situation assume different accounts of what contents visual experience has.

My defense of Thesis K goes as follows. First, I will discuss some cases in which a perceiver is disposed to recognize a K-property on the basis of visual experience. I'll argue in each sort of case that such sensitivity makes a difference to the phenomenology of visual experience. Furthermore, I'll suggest, its making a difference to visual phenomenology is a reason to think that visual experiences represent the K-property to which the subject is sensitive. The discussion will proceed with preliminary

clarifications in section 1, a discussion of why it matters whether Thesis K is true in section 2, and the case involving recognitional dispositions in section 3. I conclude in section 4 by considering some implications of Thesis K.

1. Preliminaries

Before proceeding any further, some terminology needs to be clarified: *visual experience*, its *phenomenology* and *contents*, and what it is for visual experience to *represent a property*.

Visual experiences are mental events of the sort that typically occur when a subject is seeing. These events determine the way things look to the subject. Substantive questions arise in determining what the relevant meaning of 'look' is. We cannot discern which aspects of experience are the visual ones simply by determining which English sentences of the form *It looks to S as if . . .* are true. You could speak truly when you say 'It looks to me as if Ms. Elfenbein went to Ankara,' yet not be reporting the contents of your visual experience.[4]

Visual experiences have phenomenal character, or more simply a phenomenology. The phenomenal character of a visual experience is what it is like to have that visual experience. In general, I will say that events of sensing, such as seeing, have a sensory phenomenology. Using terminology in this way, blindsight is not a form of sensing.[5]

What it is like to have a visual experience is easy to confuse with what it is like to have the overall experience—including kinesthetic, emotional, and perhaps imaginative components—of which the sensory experience is a part. Suppose you see a golden pentagon while sitting cross-legged in a garden, feeling cheerful. 'What it is like to see the golden pentagon' could reasonably be taken to pick out either the phenomenal character of the overall experience, or the phenomenal character of the visual experience of which it is a part. What it is like to see a golden pentagon differs from what it is like to see a rocky hillside. More generally, a visual experience V counts as phenomenally the same as a visual experience V′ just in case V and V′ have the same phenomenal character. V and V′ could be phenomenally the same, even though the subject of V feels cheerful while the subject of V′ feels gloomy.

What needs clarification next is the notion that visual experiences have contents. The contents of visual experience are the sort of things that have accuracy conditions. If a visual experience has the content that there is a golden pentagon in front of one, then this content is accurate just in case there is a golden pentagon in front of one.

Some philosophers have denied that visual experiences have contents, even in this minimal sense. If a visual experience is nothing but a 'raw feel,' for instance, then it has no contents. In assuming that visual experiences have contents, I am assuming that they are not merely raw feels.[6]

When experiences have content, they represent—perhaps inaccurately—that such-and-such is the case. They represent that certain things have certain properties. For example, when you see a ripe tomato under normal circumstances, your experience represents the tomato surface as being red. In general (as I suggested at the start), when experiences represent that a thing x has property F, it is representing the property F. So visual experiences represent properties.

Thesis K says that some K-properties are sometimes represented *in visual experience*. Being represented in visual experience is one way in which properties can be represented. Some visual processes represent properties, where such representation has no associated phenomenology. In contrast, when a property is represented in experience, its being so represented has an associated phenomenology.

Now, it is a theoretical question, with many competing candidate answers, exactly what relation being represented in experience has to sensory phenomenology. The notion of being represented in experience that figures in Thesis K leaves open whether there is any explanatory relation between property-representation and sensory phenomenology, and whether either of these constitutes the other. Given what's built in to the notion of being represented in experience, Thesis K amounts to this: in whatever sense the representation of color and shape properties can have an associated sensory phenomenology, the representation of K-properties can too.

Finally, for all I've said about being represented in experience, a property can be represented in experience, even if the subject of the experience possesses no concept of that property. According to some philosophers, subjects can sensorily represent a determinate hue, even when they are not disposed to recognize

that hue on subsequent occasions.[7] Experiences of colors are supposed to be the paradigm case of non-conceptual sensory representation. I don't know of any discussion of the topic that presents a K-property as an example of a property that can figure in what have come to be called 'non-conceptual contents' of experience. The arguments for Thesis K in this chapter won't challenge the implicit assumption that K-properties, perhaps unlike some color properties, can be represented in experience only if the subject has *some* sort of disposition to recognize their instances (assuming that they have multiple instances). But this assumption will not be built into the very notion of property representation in experience. That notion is neutral on what it takes for a property to be represented in experience.

2. Why It Matters Whether Thesis K Is True

I said earlier that asking what shall count as a misperception is a way of making vivid the issue surrounding Thesis K. Why does it matter whether Thesis K is true, and what counts as a misperception? It matters for at least four reasons.

First, the problem of intentionality is sometimes posed as the problem of how it is possible for a subject to be in a contentful state. A verdict on what counts as a misperception would constrain the explanandum for the case of visual experience. That is, a verdict on what may count as a misperception places a constraint on accounts of how it is possible for there to be contentful visual experiences in the first place. If visual experience cannot represent that there are peaches on the table, then whatever makes it the case that a visual experience has the content it does had better not allow that visual experiences represent the property of being a peach.

Second, there may be general skeptical worries that get going only if the contents of visual experience turn out to be informationally impoverished. Suppose, for example, that veridical experiences could only provide information about the colors and facing surfaces of objects, and not about which facing surfaces belong to the same object, or whether or not they continue out of view. Could such visual experiences play the justificatory role claimed for them by a correct theory of justification? Someone might reasonably doubt that they could. Settling what contents visual experiences have will determine whether such a challenge is worth attempting to formulate.[8]

A third reason why the truth of Thesis K matters has to do with recent research on pathological conditions known as delusions of belief, such as those found in Capgras Syndrome, which is a condition in which patients seem to believe that people close to them have been replaced by impostors. An important empirical question for psychiatry is what the nature of the delusion is: roughly, whether it is a normal response to an unusual experience, or whether instead it is an unusual response to a normal experience.[9] In forming empirically testable hypotheses about delusions of belief, it is useful to have independent support from the philosophy of perception about what sorts of contents experiences can have.

The fourth reason why it matters whether Thesis K is true relates to the role of experiences in justification. Let a w-world be a world with the actual laws of nature, in which subjects have the same perceptual equipment as we do. Consider the following claim:

> (+) If two visual experiences in a w-world differ in which properties they represent and all other factors relevant to justification are the same, then they differ in which propositions they provide justification for believing.

Suppose that visual experiences provide immediate justification for believing a proposition p, where this means that the justification provided by visual experience does not depend on any other factors. Assuming that experience provides immediate justification in virtue of the properties it represents, claim (+) will be true if any difference in properties represented makes a difference to justification provided.

Even theories that deny that there is such a thing as immediate justification can accept (+). Suppose that visual experiences provide evidential support for propositions only with the addition of certain special background beliefs on the part of the subject. According to claim (+), if two subjects in a w-world have exactly the same background beliefs (which themselves have the same epistemic status) and their visual experiences differ in what properties they represent, then different propositions will be evidentially supported by the visual experience combined with the background beliefs.

Let us take another example. Suppose that which propositions visual experiences provide justification for believing depends on the environmental conditions in which the visual experiences are had. For example, suppose that which propositions the subject is justified in believing depends on whether the belief-forming

process, of which the experience is a part, is reliable. According to claim (+), if such mechanisms in two subjects in a w-world are equally reliable and their visual experiences differ in what properties they represent, then different propositions will be such that the visual experience combined with the environmental conditions justify the subject in believing them.

If claim (+) is true, then what propositions one's visual experience contributes to providing justification for will depend on which properties visual experience represents. But is claim (+) true?

Claim (+) is very strong. It says that in any w-world, every difference in properties represented by experience matters for what an experience, combined with other factors relevant for justification, provides justification for believing. A claim at the opposite extreme says that in any w-world, no difference in properties represented by experience makes a difference for what an experience, combined with other factors relevant for justification, provides justification for believing.

This latter claim seems false. It would be odd if, in a w-world, what contents visual experiences had was totally irrelevant to what propositions the experience (together with any other epistemically relevant factors) provided justification for believing. For example, holding environmental conditions constant, compare two visual experiences, one of an undifferentiated blue expanse, the other of a dairy farm. Now consider the claim that the propositions that the experience together with these conditions provide justification for believing are exactly the same. This claim seems not to respect the basic point that what one sees makes a difference to what one is justified in believing. The falsity of this claim is enough to make the general issue of what shall count as a misperception matter. But this is compatible with the denial of (+). I'm not sure whether a claim as strong as (+) is true.[10] But I think something is true that's stronger than the basic point that what one sees makes a difference to what one is justified in believing. Consider two entirely veridical w-world experiences had by Boring and Rich. Boring and Rich are facing a fruit bowl. Boring's experience represents only colored shapes, whereas Rich's represents that there is a bowl of fruit on the table.

Now, Boring's experience supports invariances that Rich's experience doesn't. Both experiences represent properties that some rubber balls can look to have, as well as representing properties that peaches can look to have. But consider the result of combining each experience with the belief that rubber balls look to have certain shape and surface-shape (texture) properties—properties that both experiences represent. Arguably, combining this belief with Boring's experience yields some sort of evidence that there are rubber balls in the bowl: visual experience represents that there are certain colored volumes; the background belief is that some rubber balls look to have the property of being spherical and (let's say) orangey-pink.

In contrast, combining Rich's experience with the belief that rubber balls look to have certain color and shape properties does not seem to yield the same evidence. Rich's experience represents that there is fruit in the bowl (along with representing the color and shape properties that Boring's experience represents). Now, nothing is both a rubber ball and a piece of fruit. So the fact that Rich's experience represents the property of being fruit weakens the evidence for there being rubber balls in the bowl.

I think this sort of case shows that the justificatory role of experiences is not indifferent to whether it represents K-properties or not. I haven't tried to defend the claim that Boring and Rich—the subjects—are justified in believing different propositions. But I have given a reason to think that, as factors in justification, the experiences of Boring and Rich are not interchangeable.[11]

I now turn to the case for Thesis K from recognitional dispositions.

3. Recognitional Sensitivity to K-Properties

My case for Thesis K involves experiences in which the subject's beliefs about what she is seeing seem to affect visual phenomenology. Changes in beliefs about what one is seeing don't always bring about changes in one's visual phenomenology. A case in point is the Müller-Lyer lines, which continue to look as if they differ in length, even after one learns that they don't. But there seem to be other cases in which changes elsewhere in the cognitive system do bring about phenomenal changes. The argument for Thesis K depends crucially on *intuitions* about these examples, or others like it. Before turning to examples of such changes, some remarks about methodology are in order.

It is often best to avoid arguments that rest ultimately on intuition, since there can be at most a stand-off between a proponent of the

argument and someone who does not share the intuition. In this case, however, appeals to intuition of some sort are unavoidable. Perhaps this is why other defenders of Thesis K have not tried to offer arguments for it at all, but have opted instead simply to give convincing *descriptions* of the phenomenology.[12] The discussion here is an attempt to split the difference between description and argument, by starting with a minimal intuition, and then mapping out exactly what an opponent of Thesis K would have to deny if she accepts the initial intuition.

What about the initial minimal intuition? What happens if someone doesn't share it? It seems reasonable to expect there to be some intuitions that elicit broad agreement, since visual experience is something to which one has first-person access. The exact nature and limits of such access is a topic unto itself. But the intuitions on which the case for Thesis K rests are simple and modest: they are intuitions about whether there is a change in phenomenology between two sorts of situation. It would be quite a radical view that denied that there were *any* such cases in which first-person access could detect a change in phenomenology. The case for Thesis K counts on there being first-person access to the fact that there is this sort of change, but does not assume that with such access alone one can discern the exact contents of visual experience.

Let me turn to two examples of changes in the cognitive system that seem to bring about phenomenal changes. Both involve the gradual development of properly grounded recognitional dispositions.

The first example involves the disposition to recognize semantic properties of a bit of text, grounded in knowledge of how to read it. Although Thesis K concerns only visual experience, it is useful to begin with an auditory example. Almost everyone has experienced hearing others speak in a foreign language that one doesn't understand, and that one can't parse into words and sentences. The phenomenology of hearing the same speech when one does understand is markedly different.

This contrast has a visual analog. Consider a page of Cyrillic text. The way it looks to someone before and after she learns to read Russian seems to bring about a phenomenological difference in how the text looks. (Christopher Peacocke makes a similar phenomenological claim in chapter 3 of *A Study of Concepts*.)[13] When you are first learning to read the script of a language that is new to you, you have to attend to each word, and perhaps to each letter, separately. In contrast, once you can easily read it, it takes a special effort to attend to the shapes of the script separately from its semantic properties. You become disposed to attend to the semantic properties of the words in the text, and less disposed to attend visually to the orthographic ones.

The second example involves a different recognitional disposition. Suppose you have never seen a pine tree before, and are hired to cut down all the pine trees in a grove containing trees of many different sorts. Someone points out to you which trees are pine trees. Some weeks pass, and your disposition to distinguish the pine trees from the others improves. Eventually, you can spot the pine trees immediately. They become visually salient to you. Like the recognitional disposition you gain, the salience of the trees emerges gradually. Gaining this recognitional disposition is reflected in a phenomenological difference between the visual experiences you had before and after the recognitional disposition was fully developed.

The argument for Thesis K from these cases has three substantial premises, plus a premise that is unproblematic if the cases are convincing. Let E1 be the sensory experience had by a subject S who is seeing the pine trees before learning to recognize them, and let E2 be the sensory experience had by S when S sees the pine trees after learning to recognize them. E1 and E2 are sensory parts of S's overall experiences at each of these times. I'm going to call the premise that is unproblematic if the cases are convincing premise (0):

(0) The overall experience of which E1 is a part differs from the overall phenomenology of which E2 is a part.

Claim (0) is supposed to be an intuition. It is the minimal intuition one has to have, for the argument to get off the ground.

(1) If the overall experience of which E1 is a part differs from the overall phenomenology of which E2 is a part, then there is a phenomenological difference between the sensory experiences E1 and E2.
(2) If there is a phenomenological difference between the sensory experiences E1 and E2, then E1 and E2 differ in content.
(3) If there is a difference in content between E1 and E2, it is a difference with respect to K-properties represented in E1 and E2.

If no experiences represent K-properties, then there will be no difference between E1 and E2 with respect to K-properties represented in them. So if (3) and its antecedent are true, then Thesis K is too. An analogous argument could be made for the case of the Cyrillic text.

Premises (0) and (1) entail that there is a phenomenological difference between the *overall* experiences of which E1 and E2 are parts. It specifies that it is a difference in *sensory* phenomenology (the phenomenology of sensing). Premise (0), in contrast, allows that the phenomenological difference is not a difference in sensory phenomenology, but in phenomenology of some other sort.

Given premise (0), there are three ways to block the inference from these cases of recognitional dispositions to Thesis K. First, one could deny that the phenomenological changes are sensory. This would be to deny (1). Second, one could grant that they are sensory, but deny that there is any accompanying representational difference (i.e., any difference in contents of E1 and E2). This would be to deny (2). Finally, one could grant that the phenomenological changes are accompanied by a representational change, but deny that the change involves any representation of K-properties. This would be to deny (3). I will consider each of these moves in turn.

Premise (1): Non-Sensory Phenomenology?

Let me start with the first way of attempting to block the inference to Thesis K. There are various kinds of phenomenology besides sensory phenomenology. There is the phenomenology associated with imagination, with emotions, with bodily sensation, with background phenomenology, and perhaps with some non-sensory cognitive functions. If the phenomenological change described in the two cases is non-sensory, the two most plausible suggestions seem to be that it is a change in some sort of cognitive phenomenology, or in background phenomenology. Someone might be tempted to re-describe the text and tree cases so that as far as sensory phenomenology is concerned, the experiences had with and without recognitional dispositions are the same; but that the difference in phenomenology of overall experiences is due to a non-sensory factor. If these descriptions were correct, then the examples would not bear on what properties *sensory* experience represents at all, hence would not bear at all on Thesis K.

The strategy of the opponent I'm considering, then, is to re-describe the tree and text cases by invoking non-sensory phenomenology, and thereby avoid making a commitment to Thesis K. Let's consider cognitive phenomenology first.

What structure would such re-descriptions have to have? Well, first, there would have to be an *event* in the stream of consciousness, other than the event of sensing (seeing, hearing, etc.), that allegedly has the phenomenology. Just as events of sensing have an associated phenomenology, and just as events of imagining and having (some) emotions have an associated phenomenology, so too if there is cognitive phenomenology, there must, it seems, be some events in the stream of consciousness that the phenomenology attaches to.[14] Second, assuming that the mental event involves a propositional attitude of some sort, a plausible account would have to be given of the *attitude* involved in the event, and of the *content* of that attitude.[15] Finally, some reason would have to be given to think that the phenomenology involved really isn't sensory. So, for the strategy to succeed, plausible accounts are needed of four things: the *event* in the stream of consciousness that has the (alleged) non-sensory phenomenology; the mental *attitude* it involves; the *content* of that attitude; and the factors that make the phenomenology non-sensory.

The general idea behind the strategy is that the *familiarity* that one gains in gaining a recognitional disposition is reflected in cognitive phenomenology. I now want to list some of the options for the event types, attitudes, and contents that an opponent of Thesis K who followed this strategy might invoke, in accounting for this feeling of familiarity. Though the list of options is not put forward as exhaustive, they are natural ones to consider (and the only ones that come to mind after much consideration). Once they are on the table, it will be easier to assess the case against Thesis K.

It is natural to list the events and attitudes together. They include:

(i) forming a judgment;
(ii) dwelling on a belief;
(iii) entertaining a hunch or intuition;
(iv) entertaining a proposition by having it pass through your mind, without committing to its truth.

These are four sorts of events that can occur in the stream of consciousness. Entries (i)-(iii) are *commitment-involving*: the attitudes are all

related to belief, and its accompanying commitment to the truth of the thing believed. Hunches and intuitions are like beliefs in that the subject accepts their content for certain purposes. For instance, in testing a hypothesis, one may reason as if a hunch or intuition were true. Entry (iv), in contrast, does not involve any such commitment. This distinction will be useful shortly.

What about the content of the attitudes involved in the event? Since the events are supposed to be brought about in part by gaining a recognitional disposition, the contents should reflect this gain in some way. Some reasonable options include these (I'll stick to the case of the trees):

(a) *That* is a pine tree (mentally demonstrating a tree).
(b) I've seen trees with *that look* before.
(c) I recognize *that kind of tree*.
(d) *That kind of tree* is familiar.

These are supposed to be contents of mental states, rather than contents expressed by actual uses of sentences. As such, the proposal that there are attitudes that have (a)–(d) as contents involves a notion of a demonstrative thought, independent of the notion of what is expressed by an actual use of a demonstrative. The contents are analogs for thought of contents expressed by uses of sentences.

Suppose we combine any of these contents into any of the attitudes and events in the first list. Then we will have a candidate for an event with phenomenology. The denier of (1) will still owe some account, however, of what makes the event that has the phenomenology *non-sensory*. (I'm assuming that if the event is non-sensory, then so is its associated phenomenology.)

Let me now examine one instance of the strategy I've outlined for denying (1). In the tree case, the suggestion comes to this. How the tree looks before and after you become disposed to recognize pine trees is exactly the same: it looks to have certain color and shape properties. But at the moments when you recognize the tree, you experience a feeling of familiarity, and this feeling accounts for the phenomenological change before and after you gain the disposition. So, on this suggestion, the way the tree looks stays the same, before and after you become disposed to recognize it; but the phenomenology of 'taking' the tree to be familiar contributes to the phenomenal change accompanying E2. For the purpose of discussion, I'll select the event and attitude of dwelling on a belief, and content that *that kind of tree* is familiar (so, (ii) and (d)).

I'm going to raise two objections to the view that the phenomenological change in the tree case consists exclusively in a change in cognitive phenomenology, where the cognitive phenomenology is had by event and attitude (ii) with content (d). The first objection would also apply, if the event and attitude were (i) or (iii), and if it had any of the four contents listed. My second objection is more general: it would apply to any combination of the events, attitudes, and contents listed.

The first objection focuses on the events with commitment-involving attitudes. Suppose that you're an expert pine-spotter looking at some pine trees in the forest. Then someone tells you that the forest has been replaced by an elaborate hologram, causing you to cease to dwell on the belief that you're looking at a familiar tree. If an event such as (ii)(d) were what contributed to the phenomenological change before and after acquiring the disposition to recognize pine trees, then we would expect your acceptance of the hologram story to make the hologram look as the forest looked to you before you knew how to recognize pine trees. But intuitively, the hologram could look exactly the same as the forest looked to you after you became an expert. So the familiarity with pine trees does not seem to have its phenomenological effects at the level of belief.

The case against the proposal that the feeling of familiarity is conferred by a belief holds equally well against the proposal that substitutes any of the commitment-involving events/attitudes for the one I chose for purposes of discussion. Hunches and intuitions, like beliefs, seem to be attitudes that one could lose by accepting the testimony described above. If anything, hunches and intuitions are *less* resistant than beliefs are to counter-evidence—if the belief wouldn't survive accepting hologram testimony, then neither would hunches or intuitions.

The objection I've just made would not threaten a version of the strategy that invoked a non-commitment-involving attitude, such as entertaining a proposition without committing to its truth. So let us focus on a version of the proposal that appeals to an event of this sort. On this version of the proposal, in the tree case, when you look at the tree after having gained the recognitional disposition, you undergo a mental event, distinct from sensing, that has

a phenomenology of its own. This is an event (we're supposing) of entertaining the proposition that *That kind of tree* is familiar, where this proposition passes through your mind, without your committing to its truth.

Here it is important to keep in view the aspect of the proposal that posits an event (supposedly a 'cognitive' event) occurring in the stream of consciousness. This proposal predicts that there will be a phenomenological difference between your experiences of seeing the pine tree before and after you learn to recognize trees, only to the extent that such an event is occurring. If no such event is occurring, then, this proposal predicts, there will be no phenomenological change of the sort invoked in the original example.

The second objection targets this aspect of the proposal. An event's occurring in the stream of consciousness is not akin to having a tacit recognition (or misrecognition) of something as a tree. It is something explicit, rather than tacit. But the phenomenological change in the original tree example seems to be the sort that does not always involve an explicit entertaining of a proposition such as (d). Consider a comparable thought from Charles Siewert:

> [t]hink of how individual people look different to you after you have gotten to know them than they did when you first met. Notice how different your neighborhood looks to you now that you have lived there for a while, than it did on the day you first arrived. (1998, pp. 257–8)

What can happen with a neighborhood, it seems, can happen with trees as well. The phenomenological change is the sort that we can infer by remembering how different things looked before we became familiar with them. Becoming aware of the phenomenon involves thinking of something—a person, a neighborhood, or a kind of object, such as a tree—as familiar. But simply undergoing the phenomenon does not have to involve this. There need not be, it seems, an extra event, beyond sensing, for the phenomenological change to take effect.

I've raised this objection against the proposal that invokes a non-commitment-involving attitude. But it works equally well, if it works at all, against the proposal invoking a commitment-involving attitude.

At this point, the denier of premise (1) might reply to the objection by claiming that the event in question could have a content such as (d) explicitly, without the event being the sort I've described. After all, the denier might point out, sensory experience has its content explicitly, without involving something analogous to *saying* to oneself something like (d) (e.g., 'well how about that, that's a tree,' etc.).

If the putatively non-sensory event does not involve something analogous to saying to oneself something like (d), and if it is supposed to be something other than an event of visually appearing, then it becomes less clear that it is accompanying an event in the stream of consciousness at all.

Let me now consider the proposal that the phenomenal change is a change in background phenomenology, rather than a change in cognitive phenomenology attached to an occurrent event. Someone who denied premise (1) might claim that although the phenomenological difference between E1 and E2 is not sensory, neither does it belong to a specific cognitive event in the stream of consciousness.

Drunkenness and depression may be two examples of standing, background states that affect overall phenomenology. As against premise (1), someone might claim that recognitional dispositions are like drunkenness and depression in the crucial respect: they too are standing states of a subject that can affect overall phenomenology—and, indeed, the objector will claim, that is just what happens in the text and the tree cases.

To defeat premise (1) in this way, what the objector would need is a reason to think that changes in standing states can affect overall phenomenology in some way other than by causing changes in sensory phenomenology. Depression and drunkenness may involve at least some such changes: depression can cause things to look grey; drunkenness can cause them to look blurry. The relevant analogy has to be between changes in overall phenomenology that are not the result of changes in sensory phenomenology. The changes must be akin to changes in mood.

Having a recognitional disposition, however, is not phenomenologically like being in a mood at all. Moods have relatively non-local effects on phenomenology: almost *nothing* seems exciting during depression; nearly *everything* seems exciting during drunkenness. In contrast, being disposed to recognize pine trees does not have such overall phenomenological effects. So, whatever phenomenal change results from gaining recognitional dispositions, it does not seem to be a change in background phenomenology.

Let me now consider how the text example fares, if (1) is false and the phenomenological difference in how text looks before and after one learns to read it is a difference in cognitive, as opposed to sensory phenomenology.

A fan of premise (1) can grant that there are some cases in which reading a text does involve undergoing events that have a phenomenology, and that are arguably non-sensory. Lingering on a sentence while deliberating about whether it is true has a phenomenology, and arguably such an event is non-sensory. It could happen, for instance, if you weren't perceiving anything at all, but simply entertaining the proposition expressed by the sentence.

Contrast this phenomenology with that of being bombarded by pictures and captions on billboards along the highway. This seems a visual analog of the blare of a loud television, or a fellow passenger's inane cell-phone conversation. Understanding the text on the billboard as you drive by isn't a deliberate affair; rather (if the billboards have been positioned correctly), it just happens. It would please the advertisers if you lingered over every billboard's message, but no such event need occur in order for you to have 'taken in' the semantic properties of the text as you whizz by. This suggests that the 'taking in' can be merely sensory.

So far, I've considered two ways to deny (1). One way is to propose that the phenomenal change is a change in cognitive phenomenology that is attached to a specific event in the stream of consciousness. The other way is to propose that the phenomenal change is a change in background phenomenology. I've given reasons to think neither of these ways of denying premise (1) in the argument in for K will succeed. I now turn to the strategy of denying premise (2).

Premise (2): A Non-Representational Phenomenological Change?

The denier of (2) tries to block the inference from the examples of phenomenological change to Thesis K, by claiming that phenomenological changes are unaccompanied by any representational change at all.

Premise (2) is a consequence of a more general claim, one that is controversial in the philosophy of mind. This is the claim that with *any* change in the sensory phenomenology, there is a change in the content of sensory experience. But premise (2) itself is much more limited. It just makes a claim about phenomenology of the sort at issue in the two cases.

If (2) is false, then there is such a thing as a non-representational feeling of familiarity. This could be part of sensory experience, or part of some sort of cognitive event. Either way, it would be a feeling of familiarity that could be had even in the absence of perceiving, or seeming to perceive, anything as being familiar.[16] It would not represent anything as being familiar, but rather would be akin to a sensory affliction. It would be a raw feel. The proposal is not that there is merely a non-representational *aspect* to a representation of familiarity. Rather, the proposal is that the feeling of familiarity is entirely non-representational.[17]

Against this idea, my defense of (2) is that familiarity is not the sort of thing that could be felt without any representation of something as familiar. The best attempt to make the case for the contrary ends up positing a representation of familiarity after all.

One would expect a raw feeling of familiarity, if there was such a thing, to leave one with a sense of confusion, since if it was clear to the subject what was being felt to be familiar, then this would seem to make the feeling representational after all. Suppose, for example, you see someone who acts toward you as a stranger would, and this seems inappropriate to you, but you can't at first figure out why. In response to this feeling of strangeness, you might think to ask the person whether you have met before. But the feeling you have that leads you to ask it, someone might suggest, is a raw feeling of familiarity. It is a variety of 'déjà-vu.'

In the case above, the sense of confusion comes from the fact that though you take the person to be familiar, you don't recognize who they are. There are two aspects to this experience: you represent something as familiar without recognizing it, and you represent something as familiar, without at first realizing that it is so represented. The first aspect is definitive of déjà-vu: a place, or a sound, or a situation strikes you as familiar, without your being able to discern what is familiar about it. This is simply a less specific representation of familiarity—it is not a case of a feeling that does not represent anything as familiar. So the putative case of a raw feeling of familiarity does not illustrate this after all. And if that case doesn't illustrate it, it is hard to see what kind of case would.

Premise (3): Exclusively Non-K Representation?

I now consider the third response to the argument for Thesis K, which is to reject premise

(3) while granting (0)–(2). Premise (3) says that the difference in content between E1 and E2 is a difference with respect to K-properties represented in E1 and E2.

Both the tree and the text examples involve a gain in recognitional dispositions, and it will be useful to keep in mind what sort of structure recognition has. A perceiver who can recognize trees by sight seems to have some sort of memory representation, and some sort of perceptual input, such that the input 'matches' the memory representation, and the cognitive system of the perceiver registers that this is so. Empirical theories of object recognition are supposed to explain the nature of each of these components (the memory, the input, and the matching), and the mechanisms that underlie them. Part of what's at issue in the debate about Thesis K is whether visual experience is an input to such processes of recognition, or an output of such processes. Whichever empirical and philosophical theories turn out to be correct, some structure such as this seems built in to the very notion of recognition.

One sort of proposal about the contents of E2 that a denier of premise (3) might invoke would involve the notion of a pine-tree shape-gestalt. Suppose that when you learn to recognize pine trees by sight, your experience comes to represent a complex of shapes—leaf shapes, trunk shapes, branch shapes, and overall pine tree shapes. This complex is an overall pine-tree gestalt. The pine-tree-shape gestalt is general enough that it can be shared by differently looking pine trees. But it is specific enough to capture the look shared by exemplary pine trees. The pine-tree-shape gestalt is invariant across differences in shape of particular pine trees.

For an experience of seeing a tree to represent a pine-tree-shape gestalt, it need not be part of the content of experience that the tree seen is similar to other trees with respect to such-and-such shapes. It is enough simply to represent the respects in which various pine trees are in fact similar. A pine-tree-shape gestalt, then, is not by definition something that can be represented in experience, only if the subject is disposed to believe that the different things instantiating it are the same shape. But all things that have it have a complex shape property in common.

It seems plausible to suppose that pine trees share a pine-tree-shape gestalt, to the extent that pine trees, varied though they may be in size and other features, have some quite general shape properties in common. If there were such a thing as a tree-shape gestalt, then the denier of (3) could invoke this as the non-K property that E2 represents and E1 doesn't. I'm going to call this proposal for denying (3) Anti-K.

> Anti-K: E1 and E2 differ with respect to the pine-tree-shape-gestalt properties they represent, and neither represents any K-properties.

In the tree case, as Anti-K would describe it, the perceiver's experiences come to represent the tree-shape-gestalt as part of the same process by which the perceiver comes to have a memory representation 'matching' that shape gestalt.

I don't know of a knock-down argument against Anti-K. But the strategy of invoking the representation invariant color-shape complexes to underpin phenomenological changes does not seem generally available. Consider, for example, the property someone's face can have of expressing doubt. One could learn to recognize when the face of someone, call him X, was expressing doubt. X might even belong to a group of people whose faces all express the doubt in the same way. Initially, one might not know that X and his kin are expressing doubt when they look that way. But this is something one could learn to recognize by observing them. In this sort of case, it seems implausible to suppose that there must be a change in which color and shape properties are represented before and after one learns that it is doubt that the face so contorted expresses. One could initially wonder what the contortion of the face meant, and come to believe that it is an expression of doubt only after repeated sightings of it and interaction with the person. This change in interpretation seems to be one that could be accompanied by a phenomenological change as well.

Once they are adjusted to be about the face case, the other two premises of the argument still seem to go through. Exactly the same considerations apply in the case of premise (2). In premise (1), the argument for ruling out non-sensory phenomenology also seems to go through as before, but in the face case another alternative to sensory phenomenology seems relevant—namely, emotional phenomenology. Here, it seems possible in principle that X could learn to detect a look of doubt on Y's face without X's having any emotional response—Y might not be anyone significant for X, such as a talking head on television.

As for premise (3), an opponent who granted the initial intuition that there is some phenomenal change accompanying the gain of a recognitional disposition might say that the

phenomenal change is sensory, but that the novel phenomenology is associated merely with coming to represent the property of being a familiar expression. This option seems to be ruled out by considering a variant of the face case involving two subjects. Consider a counterfactual situation in which X contorts his face in exactly the same way, but in which that contortion expresses bemusement rather than doubt. One could come to learn that it expresses bemusement in the same way as in the first case, by extended observation and interaction. But it seems plausible to suppose that the phenomenal change in each case would be different: one sort of phenomenology for recognizing the doubtful expression, and another sort for recognizing bemusement.

Finally, return to the text example to see how it fares with respect to premise (3). The original intuition was that before and after you learn to read Russian, the same page of Cyrillic text will look differently to you. You might love the look of Cyrillic script, keep a page nearby at all times, and study its shapes carefully. Then, after learning to read Russian, you see by reading it that it is a page of insults. Even if you attended to colors and shapes of the Cyrillic script as thoroughly as possible before learning to read it, you would still experience the page differently once it became intelligible to you.

I've argued that gaining a disposition to recognize K-properties can make a difference to visual phenomenology, and that this difference is accompanied by a representation of K-properties in visual experience. In the next section, I consider why it matters whether Thesis K is true.

4. Some Implications of Thesis K

I will conclude by discussing some implications of Thesis K, and of the considerations I've appealed to in support of it.

First, if Thesis K is true for reasons of the sort given here, then visual perception as a whole is at most partly informationally encapsulated: which contents visual experiences have can be influenced by other cognitive processing. Even if visual experiential representations of some properties cannot be influenced by what happens in other parts of the cognitive system, visual experiential representations of other properties can be. If the argument here is sound, our perceptual systems may include modular 'input systems' of the sort described by Fodor 1983, but these systems will not be ones with which visual *phenomenology* is exclusively associated.

Second, my case for Thesis K has proceeded without appealing to any specific theory of intentionality for experiences. The pine-tree example might be taken to suggest that one of the K-properties that visual experiences can represent is the natural kind property of being a pine tree. Since it is widely held that any mental state that represents natural kind properties has contents that are externally determined, it is natural to ask what relation Thesis K bears to the thesis that some experiential contents are externally determined.

Externalism about experience content is the view that physical duplicates can differ in which contents their experiences have. Since Thesis K is silent on what makes it the case that experiences represent what they do, it is clearly compatible with externalism about experience content.

Suppose that Thesis K is made true by visual experience representing natural kind properties, such as the property of being a pine tree. And suppose one accepts externalism. It is open to someone who accepts both of these claims to hold that physical duplicates whose environments differ (where only one includes pine trees) have the same visual phenomenology. This would entail that the property of having that visual phenomenology is not *identical* with the property of representing the property of being a pine tree in experience.[18] But it is compatible with the view that that visual phenomenology *supervenes* on the contents of visual experience.

Thesis K is also compatible with the denial of externalism. Even if one accepts that natural kind properties can be represented in visual experience only if externalism about experience content holds, one need not accept such externalism in order to accept Thesis K, because Thesis K can be true even if natural kind properties are not represented in experience.

Consider the case discussed earlier involving dispositions to recognize pine trees. If one rejects externalism but accepts that E2 in the example is an experience that represents some K-property, one has two options. First, one can hold (contrary to the received view) that the property of being a pine tree can be represented even by someone who was never in contact with pine trees. Second, one can hold that the K-property that comes to be represented in E2 is not the property of being a

pine tree, but a more general K-property (possibly a kind property that is not a natural kind property) that both pine trees and superficially similar trees share.

In sum, although Thesis K is compatible with externalism about the contents of experience, it does not require it. The argument for Thesis K does not appeal to any theses about content-determination. Depending on which such theses one accepts, however, and depending on views about the exact relation between content and phenomenology, one may reach different verdicts on exactly what K-property would come to be represented in the pine-tree example (assuming that the rest of the argument for Thesis K is accepted).

Finally, if Thesis K is true, then it seems reasonable to expect that K-properties other than the property of being a pine tree (or some more general K-property) and semantic properties of texts are represented in visual experience. There are two routes to generalizing the conclusion beyond the two specific properties used as examples in the argument. One route is to run exactly analogous arguments for other cases in which becoming sensitive to property instances has an effect on overall phenomenology. The argument structure leaves open, however, whether analogous considerations will always be available—and to that extent, the first route to generalizing the conclusion may turn out to be somewhat limited.

There is, however, a second route to generalizing the conclusion. Thesis K has some prima facie plausibility. One role of the argument given here is to provide positive reason to think that at least some K-properties are represented in visual experience. And if some such properties are, then it is plausible to think that others are as well.

REFERENCES

Austin, John L. *Sense and Sensibilia* (Oxford: Clarendon Press, 1962).

Block, Ned, "Mental Paint and Mental Latex," in *Philosophical Issues 7: Perception*, Enrique Villanueva, ed., (Atascadero, CA: Ridgeview Publishing Co., 1996), pp. 19–49.

Byrne, Alex, and Hilbert, David, eds., *Readings on Color*, vol. 1 (Cambridge, MA: MIT Press, 1997).

Coltheart, Max, and Davies, Martin, eds., *Pathologies of Belief* (Oxford: Blackwell, 2000).

Fodor, Jerry. *The Modularity of Mind* (Cambridge, MA: MIT Press, 1983).

Gunther, York H., ed., *Essays in Non-Conceptual Content* (Cambridge, MA: MIT Press, 2003).

Maher, Brendan A., "Delusional Thinking and Perceptual Disorder," *Journal of Individual Psychology* 30 (1974): pp. 98–113.

_____ "Anomalous Experience in Everyday Life: Its Significance for Psychopathology," *The Monist* 82 (1999): pp. 547–70.

Martin, Michael, "The Reality of Appearances," in *Thought and Ontology*, Mark Sainsbury, ed., (Milan: Franco Angeli, 1997), pp. 81–106.

Peacocke, Christopher. *A Study of Concepts* (Cambridge, MA: MIT Press, 1992).

Shoemaker, Sydney, "Phenomenal Character," *Noûs* 28 (1994): pp. 21–39; reprinted in Byrne and Hilbert (1997), pp. 227–46.

Siewert, Charles P. *The Significance of Consciousness* (Princeton, NJ: Princeton University Press, 1998).

Thau, Michael. *Consciousness and Cognition* (New York: Oxford University Press, 2002).

Tye, Michael. *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995).

_____ *Color, Consciousness and Content* (Cambridge, MA: MIT Press, 2000).

Young, Andrew W., "Covert Face Recognition in Prosopagnosia," in Andrew W. Young, *Face and Mind* (Oxford: Oxford University Press, 1998), pp. 282–312.

NOTES

Versions of this paper were presented at Vassar, Utah, the 2002 NEH Summer Institute on Consciousness and Intentionality, Syracuse, Arizona, Michigan, ANU, Taipei, University of North Carolina-Chapel Hill, and New York University. I am grateful to the audiences on these occasions for much helpful discussion. For criticism and discussion along the way, many thanks to Ned Block, Alex Byrne, David Chalmers, Fred Dretske, Rich Feldman, Hartry Field, Justin Fisher, Michael Glanzberg, Güven Güzeldere, Terry Horgan, Amy Kind, Jeff King, Bill Lycan, John Morrison, Ram Neta, Christopher Peacocke, Richard Price, Jim Pryor, Stephen Schiffer, Charles Siewert, Maja Spener, Jason Stanley, Daniel Stoljar, Scott Sturgeon, Jonathan Vogel, and the editors of this volume.

1. There has been much discussion recently about which properties in the natural world (if any) are the colors, and about whether any of those properties are the same as the properties represented in color experiences (see essays in Byrne and Hilbert 1997). Though hardly anyone denies that colors are

represented in color experience (though see Thau 2002), some have proposed that visual experience represents properties easily mistaken for the colors. (Defenders of this last view include Shoemaker 1994, and this volume, chap. 13.) Neither of these positions departs very far from the intuition that properties very much like colors are paradigms of what is presented in visual experience.

2. K-properties also exclude some properties that one might think of as kinds, such as the property of being red.
3. Though I have assumed here that experiences have contents, in the sense introduced in the text, the main question of the chapter can be posed without assuming that they do. Experience may represent properties, even if it consists in a relation to a perceived particular object and its property instances, as some disjunctivists about visual perception hold. (see Martin 1997). An analogous question would then arise: namely, which properties are such that their property instances can (partially) constitute experiences? Similarly, if experience is the having of sense-data, where these are sensory afflictions that are not assessable for accuracy, the analogous question is what properties sense-data can have.
4. One might question whether there is any use of 'looks' that is appropriate for this stipulation. As J. L. Austin 1962, p. 43, pointed out, gasoline looks like water. This seems to be a fact about gasoline that obtains independently of anyone's mental states. (We seem to be able to make sense of the idea that gasoline would look just the same, even if there were no perceivers.) Inspired by Austin, one might conclude that there is no mental property we have while seeing, in virtue of which things look the way they do; and, therefore, the working definition offered is a non-starter, since it says that visual experience is just such a mental property.

   The Austin-inspired point brings out that there are multiple uses of 'looks.' Even if its use in 'gasoline looks like water' does not tell us about any particular perceiver's mental state, there are other uses of 'looks' that do tell us about this, as when we say 'it looks to S as if there is something red and white over there.' It could look this way to S even if there is nothing red and white over there, whereas (worries about fiction aside) 'gasoline looks like water' could not be true if there were no such thing as gasoline. This is relevant use of 'looks.' Since 'looks' has such a use, our working definition of visual experience is not doomed from the start: the Austinian use of 'looks' is not the only use available.
5. In there does not seem to the subject to be anything in her visual field, if forced to guess between certain parameters (e.g., which way a line is oriented), subjects guess correctly more than half the time. For further discussion of perception without awareness see Dretske (this volume, chap. 4).
6. This leaves open that visual experiences can also have intrinsic, non-representational features of some sort, where these are not themselves truth-apt, and are also not parts of contents. So I am not assuming anything about the existence of such non-representational features, one way or the other.
7. For discussion, see the articles in Gunther 2003.
8. Even if Thesis K is true, that would not necessarily be the end of the skeptical challenge mentioned. In principle, merely entertaining the skeptical could change the contents and phenomenology of an experience in such a way that even if the experience started out representing K properties, as a result of entertaining the skeptical hypothesis it ceased to represent them. It is an open question whether entertaining the skeptical hypothesis would change experiences in just this way.
9. For discussion, see Coltheart and Davies 2000. More generally, Davies, Coltheart, and colleagues agree with Maher 1974, 1999, that several delusions in addition to Capgras are beliefs, and that part of what makes them pathologies is an experiential component in response to which the subject forms the delusional belief. Given these assumptions, one task for future research is to develop and test hypotheses about the nature of the experience.
10. One might propose to accept (+) on the basis of the claim that one is justified in ascribing to oneself an experience that represents properties F and G if one has an experience that represents F and G and not otherwise. I'm dubious about the latter claim, however.
11. I've been considering only one type of factor besides experience that is relevant to justification—namely, background beliefs. But I suspect you could make a similar argument using different non-experiential factors, such as external ones.
12. A contemporary example of such a writer is Charles Siewert, who offers excellent descriptions of the phenomenon (1998, chap. 7).
13. Peacocke writes:

    > Once a thinker has acquired a perceptually individuated concept, his possession of that concept can causally influence what contents his experiences possess. If this were not so, we would be unable to account for differences which manifestly exist. One such difference, for example, is that between the experience of a perceiver completely unfamiliar with Cyrillic script seeing a sentence in that script and the experience of one who understands a language written in that script. These two perceivers see the same shapes at the same positions. . . . The experiences differ in that the second perceiver recognizes the symbols as of particular orthographic kinds, and sequences of the symbols as of particular semantic kinds. (1992, p. 89)

    In this passage, in addition to claiming that two experiences of reading Cyrillic text would differ phenomenally, Peacocke also seems to be making an argument, with the phenomenal claim as a premise, that which concepts one possesses can causally influence which contents one's experiences have. I'm endorsing the phenomenal premise, without endorsing the argument.
14. To say that an event of sensing has an associated phenomenology leaves open whether there is any phenomenological commonality to all such events. This seems implausible for sensing as such, though perhaps more plausible within each of the modalities. But the demand on the denier of (1) is merely to show that there is an event that is not sensory and has an associated phenomenology, not to show that there is a phenomenological common element to all events, or to all events involving the same attitude, that are both cognitive and have an associated phenomenology.
15. When I discuss the strategy of denying (2), I will consider a version of this strategy that allows

cognitive phenomenology to be had by an event that does not involve any propositional attitudes, or contents thereof.

16. One might think that recent findings about the structure of face-recognition provides evidence for the existence of a non-representational feeling of familiarity. According to some neuropsychologists, the face-recognition system has at least two components: an affective component that registers when a face is familiar, and a semantic component devoted to recognizing faces (see Young 1998). These elements seem to come apart in prosopagnosics, who have the same differential affective reactions as normal perceivers do to pictures of familiar famous people, on the one hand, and to pictures of complete strangers, on the other (as measured by skin conductance tests), but who claim not to know who any of the people pictured are. Even if there is a mechanism devoted to affect of familiarity, however, that does not show that there is non-representational *phenomenology* of familiarity. The structure of underlying mechanisms of face recognition may not be mirrored by phenomenology. For us to have the phenomenology of seeing a familiar face, it may be that more than positive skin-conductance reaction is needed. (Indeed, it seems that more is needed, since otherwise we would expect the prosopagnosics' reports to be more equivocal than they are, to the effect that the person pictured seems familiar, yet cannot be named).
17. Contrast the case of color, where some philosophers argue that there are both non-representational and representational features of color experience: color properties are represented in experience, but color experiences also have non-representational features (e.g., Block 1996).
18. For a defense of this position, see Tye 1995, 2000.

5 Self-Knowledge and Other Minds

Part I of this book, and a large amount of parts II–IV, focused on the *metaphysics of mind*: what is the nature of the mind and of mental states? In recent years, there has been increasing attention to the *epistemology of mind*: how do we know about the mind and mental states? The epistemology of mind divides into two main topics. There is the problem of *self-knowledge*: how do we know about our own minds and mental states? There is also the problem of *other minds*: how do we know about others' minds and mental states?

The first four chapters deal with the problem of self-knowledge. Fred Dretske (chapter 59) lays out one version of the problem by asking 'How do you know you're not a zombie?' That is, how do you know introspectively that you have conscious mental states at all? The problem as Dretske sees it arises from the fact that in ordinary perception and thought, we are aware of the objects that we are perceiving or thinking about, but we are not awareness of our awareness itself. (This is a version of the transparency thesis discussed by Tye in chapter 27.) Without this, how can we know introspective that we are perceiving or thinking about these things at all? Dretske canvases a few possible solutions but is dissatisfied with all of them, and ends up wondering whether we really do know we are conscious at all.

Alex Byrne (chapter 60) offers a solution to Dretske's problem. He gives an account of self-knowledge that is consistent with the transparency of mental states. Byrne starts with an observation by Gareth Evans, that in making a self-ascription of belief one's "eyes are directed outward" at the world. Byrne argues that introspection starts with awareness of objects and properties in the world, and then deploys special transparent "epistemic rules" to arrive at beliefs about our mental states. He suggests that the same framework can be generalized to explain our knowledge of our desires and other mental states.

Eric Schwitzgebel (chapter 61) argues that our self-knowledge is much poorer than we ordinarily think. Philosophers often suggest that introspective knowledge is more secure than knowledge of the external world, but Schwitzgebel argues that things are reversed. He argues that ordinary introspection is highly unreliable. We have poor knowledge of our emotional lives, of our own perceptual experience, and of our thoughts. One source of this conclusion is that theorists often differ greatly about the extent these things: for example, whether there is an experience of thinking. So it is plausible that many people are getting this wrong. Perhaps introspection is reliable in certain narrow domains, or for a narrow range of thinkers, but in the ordinary case our self-knowledge is poor.

L. A. Paul (chapter 62) argues that we often lack knowledge of our future selves, and this causes problems for making decisions about the future. She focuses on the case of

deciding whether to have children. Drawing on Jackson's case of Mary in the black-and-white room (chapter 62), she argues that we cannot know in advance what it will be like to have children. Having a child is *epistemically transformative* in giving us new knowledge. Paul argues more strongly that having a child can be *personally transformative* in changing one's fundamental values and preferences. She argues that because of these transformations, it is impossible to make a fully rational decision in advance about these things.

How can we know whether others have minds? In everyday life, we accept that others have minds like ours, and this acceptance seems unproblematic. But philosophically, it is not obvious how the belief in other minds is grounded. How do we know that others are not mindless zombies? Bertrand Russell (chapter 63) argues that our belief in other minds is grounded in an *analogy* with our own case: roughly, that others are broadly similar to ourselves, that we have minds, so that others have minds. This raises many questions (should one accept an analogy based only on a single case?), but it is far from obvious what the alternatives are.

In chapter 64, Joshua Knobe and Jesse Prinz apply the methods of *experimental philosophy* to the problem of other minds. Experimental philosophy studies people's philosophical judgments and intuitions, typically by asking them philosophical questions. Knobe and Prinz asked people a range of questions about which other entities have minds. They focused especially on whether group entities such as a corporation can have mental states. Can Microsoft believe something? Can it feel pain? They found that people are much more willing to ascribe beliefs and desires than states like feeling pain or feeling depressing. They suggest that people are much less willing to ascribe phenomenal consciousness to group entities than they are to ascribe states without phenomenal consciousness. They conclude that the 'folk psychology' of other minds is less functionalist and less tied to behavior than often thought.

A crucial aspect of the problem of other minds is the problem of *animal minds*. Do non-human animals have minds, and if so which, and how do we know? We have encountered this problem at least twice already: T. H. Huxley (chapter 6) discussed the Cartesian idea that non-human animals lack minds completely, and Thomas Nagel (chapter 24) suggested that we can never completely know the mind of a bat. In chapter 65, Peter Godfrey-Smith focuses on the case of the octopus. He argues that although octopuses are very different from humans, we can know a good deal about what it is like to be an octopus.

FURTHER READING

The transparency view of self-knowledge was introduced by Gareth Evans 1982. Many articles on introspection are collected by Smithies and Stoljar 2012. Byrne 2018, Paul 2016, and Godfrey-Smith 2016 spell out their ideas at book length. Avramides 2001 overviews the problem of other minds. Knobe and Nichols 2013 is a collection of articles on experimental philosophy with a number of articles in experimental philosophy of mind.

Avramides, A. *Other Minds* (London: Routledge, 2001).

Byrne, A. *Transparency and Self-Knowledge* (Oxford: Oxford University Press, 2018).

Evans, G. *The Varieties of Reference* (Oxford: Oxford University Press, 1982).

Godfrey-Smith, P. *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness* (New York: Farrar, Straus, and Giroux, 2016).

Knobe, J., and Nichols, S. *Experimental Philosophy*, vol. 2 (New York: Oxford University Press, 2013).

Paul, L. A. *Transformative Experience* (Oxford: Oxford University Press, 2014).

Smithies, D., and Stoljar, D. *Introspection and Consciousness* (Oxford: Oxford University Press, 2012).

How Do You Know You Are Not a Zombie?

Fred Dretske

I'm not asking *whether* you know you are not a zombie. Of course you do. I'm asking *how* you know it. The answer to that question is not so obvious. Indeed, it is hard to see how you *can* know it. Wittgenstein 1921, 1961, p. 57, didn't think he saw anything that allowed him to infer he saw it. The problem is more serious. There is nothing you are aware of, external or internal, that tells you that, unlike a zombie, you are aware of it. Or, indeed, aware of anything at all.

1. Veridical Perception

To better understand the problem assume—what you probably already believe anyway—that when you open your eyes, in perfectly normal conditions, you see physical objects: people, trees and houses. Your awareness of these objects is not mediated by more direct awareness of mental particulars—for example, sense data, impressions, ideas. Watching your son do somersaults in the living room is not like watching the Olympics on television. Perception of your son may involve mental representations, but, if it does, the perception is not secured, as it is with objects seen on television, *by* awareness of these intermediate representations. It is the occurrence of (appropriately situated) representations in us, not our awareness of them that makes us aware of the external object being represented. Call this *direct realism*, and assume, for the moment, that it is true.

In normal (that is, veridical) perception, then, the objects you are aware of are objective, mind-independent objects. They exist whether or not you experience them. It follows that the properties you experience when you perceive these objects, if they are properties of anything, are properties of mind-independent objects. The properties you are aware of are properties of—what else?—the objects you are aware of. The conditions and events you are conscious of—that is, objects having and changing their properties—are, therefore, completely objective. They would be the same if you weren't aware of them. Everything you are aware of would be the same if you were a zombie.[1] In having perceptual experience, then, nothing distinguishes your world, the world you experience, from a zombie's. This being so, what is it about this world that tells you that, unlike a zombie, you experience it? What is it you are aware of that indicates you are aware of it?

Perceptual experiences (we hope) carry information about what you are aware of but this is quite different from carrying the information that you are aware of it. It is important to keep this distinction—the distinction between the *content* of awareness and the *awareness* of content—in mind when thinking about how we know we are conscious. As Burge 1988, 1996 and Heil 1988 argue, knowing what you think is easy because the content of the lower-order thought—that, for example, there is beer in the fridge—is automatically embedded in the higher-order thought whose content is that you think there is beer in the fridge. So if you are not a zombie, if you think (and think you think) at all, you don't have to worry about how you know *what* you think. You think whatever you think you think, and this is so by virtue of the fact that what determines what you think you are thinking is whatever you happen to be thinking. Our question, though, is not a question about content. It is not a question about how we know, in the case of thought, what we are thinking. It is a question about how we know we are thinking it—a question about the attitudinal aspect of thought (see Bernecker 1996, Dretske 1995). It is a question about how one gets from what one thinks—that there is beer in the fridge—to a fact about oneself—that one thinks there is beer in the fridge. What you see—beer in the fridge—doesn't tell you that you see it, and what you think—that there is beer in the fridge—doesn't tell you that you think it either.

What makes us so different from zombies are not the things (objects, facts, properties) we are aware of but our awareness of them; but this, our awareness of things, is not something we are, at least not in perceptual experience, aware of. So if you are, as you surely are, aware that you are not a zombie—aware, that is, that you are aware of things—what is it you are aware of that tells you this?

From Brie Gertler, ed., *Privileged Access: Philosophical Accounts of Self-Knowledge* (Ashgate, 2003).

Perception, broadly construed, certainly provides information about the self. As we move around there is, in vision for example, 'self specifying' information (Gibson 1979). This enables you to see where you are going (a fact about yourself) without seeing yourself. There are, in addition, proprioceptive systems that provide information about the body, information about pressure, temperature, posture, balance, fatigue, position, and so on. These sources of information about the conscious self, however, supply information about the embodied self, the vehicle of consciousness, not information about its consciousness. Zombies, after all, have bodies too. They move around. They lose their balance. A zombie's arms and legs, just like ours, occupy positions. Their muscles get fatigued (zombies are not exceptions to the laws of thermodynamics). So the conditions we receive information about in proprioception, just like the conditions we receive information about in exteroception, do not indicate that we are not zombies. These sources of information don't tell us whether we actually perceive, whether we are conscious of, those conditions of the body that we receive information about.[2]

It is true that, besides seeing objects in the world, you see these objects from a point of view. There is a perspective we have on the world, a 'boundary,' if you will, between things we see and things we don't see. And of the things we see, there are parts (surfaces) we see and parts (surfaces) we don't see. This partition determines a point of view that changes as we move around. Since zombies don't have points of view, it may be thought that this is our way of knowing we are not zombies. Although everything we see exists in the world of a zombie, what doesn't exist in the world of a zombie is this egocentric partition, this boundary, between things (and surfaces) we see and things (and surfaces) we don't see; and the fact that there is, for us, this point of view, this perspective, is what tells us we are not zombies.[3]

Points of view, perspectives, boundaries and horizons certainly exist in vision, but they are not things you see. You don't see them for the same reason you don't feel the boundaries between objects you touch and those you don't. Tactile boundaries are not tactile and visual boundaries are not visible. There is a difference between the surfaces you see and the surfaces you don't see, and this difference determines a 'point of view' on the world, but you don't see your point of view. That is not an additional object you are (visually) aware of any more than tactile boundaries are additional objects you feel. You may, of course, be aware that you have a point of view just as you are aware that there are tactile boundaries. You may know (and in this sense be aware of) this fact. But awareness of the fact that you have a point of view does not tell us how you know you have a point of view. It merely restates (using the words 'aware that') that you know it. We still haven't been told how you know it. If the things you see don't tell you, what is it that makes you aware of the fact that you see (hear, feel, etc.) some things and not others?

The fact that you can, up to a point, choose what you see (feel, hear) and when you see (feel, hear) it is of no help. Opening and closing your eyes makes a difference, it is true, but the difference it makes is not a difference in *what* you see. It's a difference in your seeing it. You are, to be sure, aware of this difference, aware that a change takes place when you open and close your eyes, but this, once again, is exactly the piece of knowledge whose source is in question. How do you know you see things when your eyes are open but not when your eyes are closed? It cannot be *what you see* that tells you because what you see when your eyes are open (for example, the morning newspaper) doesn't depend on your seeing it. The newspaper remains the same when you close your eyes.

One must be careful here to distinguish the difference between two kinds of difference, the difference between:

(a) feeling (seeing, hearing, etc.) x, and,
(b) not feeling (seeing, hearing, etc.) x.

and the difference between:

(c) feeling x (glass, say), and,
(d) feeling y (sandpaper, say).

When you feel the difference between smooth glass and rough sandpaper, this is a difference you actually feel. A tactile difference. The difference between feeling smooth glass and not feeling it is not a difference you feel. It is not a tactile difference. You can describe how you know that the texture of glass is different from the texture of sandpaper by saying you felt the difference. You come to know that they differ by actually feeling the difference. But that can't be the way you know that there is a difference between feeling glass and not feeling glass. That is a difference you can't feel. So if you know that there is, for you, a difference between feeling and not feeling things—and,

therefore, that you, unlike a zombie, sometimes feel things—we still need an answer to how you know this.

We are left, then, with a preliminary conclusion: there is nothing we perceive that tells us we are conscious. This conclusion may seem obvious—especially when we remember that it was reached by assuming perceptual realism. It is obvious once you think about it. But we have a tendency not to think about it. We have a tendency to suppose that merely being aware of objects—seeing, feeling, and smelling them—is, by itself, enough to make us aware of the fact that we are aware of them and, thus, conscious.[4] It isn't. If you know you are not a zombie, the fact that you are not a zombie, the fact that you are actually conscious of things, is not how you know it.

2. Objects We are Necessarily Aware of

Yes, but not all experience is veridical. There are illusory experiences: hallucinations, dreams and after-images.[5] Maybe the dagger I see exists, and exists pretty much as I see it, in a zombie's world, but the 'dagger' Macbeth experienced doesn't. He was aware of something that could not exist in a zombie world since if he were a zombie, what he experiences—that mental image, that 'false creation'—would not exist. Not only are there illusory experiences, there are bodily sensations. Each of us has, and each of us knows we have, headaches, itches, tickles, feelings of anger, fear, hunger and thirst. These are sensations that zombies do not have. Zombies might (depending on the kind of zombies we imagine) exhibit the symptoms of pain—they might groan, hold their head, and take aspirin—but they don't feel pain. They may also, in some dispositional sense, be thirsty (that is, exhibit a tendency to drink after periods of deprivation), but they don't feel thirsty. We do. And we know we do. That, surely, is how we know we are not zombies.

Many will find this argument convincing. So did I a year ago. But I no longer think it works. What we are looking for, remember, is a way of knowing that, unlike zombies, we are conscious of things. The argument just given shows that there are things we are aware of that zombies are not, things we feel that zombies do not, but it doesn't tell us how we know that, unlike zombies, we are actually aware of—that is, actually feel—these things. It may turn out that what we are aware of, what we feel, when we are in pain or thirsty are things of a sort that also occur in zombies. Zombies just aren't aware of them.

I will be told that pains, tickles and feelings of thirst are—by definition, if you will—mental events that are necessarily conscious. Maybe a person can be thirsty without being aware of it, but an unconscious person can't feel thirsty. Feelings are like pains. If you aren't aware of them, they do not exist. Feelings are felt, and feeling x is being conscious of x.[6] Reid 1785 and Shoemaker 1986 surely speak for many when they insist that being in pain and feeling pain are one and the same thing. It is just nonsense to talk about pains we (or zombies) are not aware of. If we (or they) aren't aware of them they aren't pains.[7]

Tickles and other bodily sensations are the same. They are mental events that we are necessarily conscious of. Anyone who has had an itch or a tickle has experienced something that can only exist in a conscious being, a being who is, in particular, conscious of the itch or tickle. *That* is how we—those of us who have itches and tickles—know we are not zombies.

But the fact—if it is a fact, and I'm willing to grant that it is—that we are necessarily conscious of our own pains and tickles, our own feelings of anger, fear, and hunger, our own imagery, doesn't help with our problem. It merely changes the way the problem is posed. If awareness of x doesn't tell you that you are aware of x–as your awareness of trees and people doesn't (see section 1)—awareness of x when x is something one is necessarily aware of isn't going to tell you either. If a pain (itch, tickle, visual image) is an object (state, event, process[8]) one is necessarily aware of, then it certainly follows that if one is in pain (has an itch or a tickle), one is not a zombie. This, though, merely invites the question: how does one know it is pain (an itch, a tickle) one feels? What is it that tells you that what you feel in your tooth is something you feel in your tooth, something you are actually aware of, and not the sort of thing that can occur, without being felt, in the tooth of a zombie?

To understand the problem here, think about *crocks*. Crocks are, by (my) stipulative definition, rocks that you (not just anyone, but you in particular) see, rocks that, you are therefore (visually) aware of. When you see a crock, then, there is something you are aware of—a crock (that is, a rock you see)—that could not exist if you were not conscious of it. There are no crocks in zombieland. There can't be.

So when you see a crock, there is something you are aware of—a crock—that depends for its existence on your being aware of it. If you closed your eyes, lost consciousness, or became a zombie, crocks would vanish.[9] To echo both Reid and Shoemaker: crocks and your awareness of crocks are one and the same thing.

So if our itches and tickles—things which cannot exist if we are zombies—are an answer to our question, so are crocks. You know you aren't a zombie because there are things, crocks, that could not exist if you were a zombie.

It should be clear, though, that crocks are not an answer to our question. So neither are itches and tickles. The reason crocks are of no help is that there is nothing about a crock that tells you it is a crock. Crocks, after all, look much the same as—in fact, they are absolutely indistinguishable from—rocks. Saying how you know you are not a zombie is just a way of saying how you tell crocks from rocks. The same is true of pains (itches, tickles, etc.). If pains are sensations we are necessarily aware of (and, of course, your pains are sensations you are necessarily aware of), then there must be something else, something that isn't pain—call it *protopain*—that has all the properties you are aware of when you experience pain except for the relational one of your being aware of it. Protopain is what you have left when you subtract your awareness of pain from pain. Protopain is to pain what rocks are to crocks. Subtract your awareness of a crock from a crock and you are left with a rock. Subtract your awareness of a pain from a pain and you are left with a protopain. And just as rocks, but not crocks, inhabit the world of a zombie, protopain, but not pain, occurs in zombies. Zombies might be full of protopain, prototickles, protohunger, and protoimagery. Lacking consciousness, though, they aren't aware of them. As a result, these internal objects do not qualify as pain, tickles, hunger and imagery. Since protopain has exactly the properties you are aware of when you feel pain (just as rocks have the properties you are aware of when you see a crock), protopains are as indistinguishable from pain as are rocks from crocks. So we come back to the question: how do you know it is pain you feel and not merely protopain?

This talk of protopain (prototickles, etc.) may sound like a philosopher's gimmick, but it is merely a way of dramatizing the difference between the act and the object of awareness, a distinction that lies at the heart of (what I regard as) a plausible theory of sensation (see, for example, Armstrong 1961, 1962; Dretske 1995; Lycan 1996; Pitcher 1971; Tye 1995). This theory regards pain (and other bodily sensations) as representations of bodily conditions. Pain, for instance, is awareness of injury, stress or irritation to some part of the body. On this view of pain, what I am calling protopain is simply the bodily condition we 'perceive,' the physical condition we are made aware of, when we are in pain. Pain is our awareness of protopain and is, therefore, necessarily conscious in the sense that pain cannot occur without consciousness (of protopain). These bodily conditions, these protopains, these internal objects of awareness, however, occur in zombies as well as in us. Zombies just aren't aware of them. They don't feel them. Neither do we when under anaesthesia.

I have had impatient people react to this argument in the following way. Look, when I am in pain I cannot (like a zombie) simply be having a protopain since if I were, I wouldn't be in pain. But I am in pain. And I know it. So I'm not a zombie.

I agree you are not a zombie. I also agree that you know this and know, therefore, that the pain you feel is genuine. It is not just protopain. But that has never been in dispute. The question, remember, is not whether you know any of this, but how you know it. If there is nothing that distinguishes these objects of awareness, your pain from your protopain, except the associated (with pain) act of awareness, there is nothing you are aware of when you feel pain that tells you it is pain you feel and not merely protopain. That is the problem, and mere insistence that you know you are in pain isn't an answer to how you know it. I also agree that there are things that exist, and you know they exist, in your garden—namely, crocks—that couldn't exist if you were a zombie, but that doesn't tell me how you know you are not a zombie. Not unless you can tell me how you know there are crocks in your garden.

Taking a cue from the discussion above (where pain was identified with an awareness of protopain), it may be objected that my rock/crock analogy misrepresents the phenomena of bodily sensations. It isn't that pain stands to protopain as crocks stand to rocks. Pain isn't the thing we are aware of when we are in pain (as crocks are the rocks we are aware of when we see a rock). A better analogy would compare pain not to the object of awareness, the thing we are aware of when we are in pain, but, rather, to the act of awareness. Pain is our awareness of

something (a bodily condition?) an awareness of which is the pain. On this picture of things, the analogy looks more like this: pain stands to protopain (whatever we are aware of when we are in pain) in the way our awareness of rocks stands to rocks. Just as it is easy to distinguish our awareness of rocks from rocks (unlike a crock, an awareness of a rock doesn't look at all like a rock) it is easy to distinguish pains (awarenesses of protopain) from the protopains that exist in zombies. If we adopt this model, then, being aware of pain is awareness of something (an awareness of protopain) that cannot exist in a zombie. It not only cannot exist in a zombie, it is easily distinguished from the protopains that might occur in zombies. Maybe *that* is how we know we are not zombies: we are aware of our own awarenesses of things.

3. Awareness of Awarenesses

If objects one is necessarily aware *of* (for example, crocks) isn't the answer, then maybe objects that are themselves awarenesses (of other things—for example, rocks) will do the trick. We see things. Zombies don't. We have experiences of things. Zombies don't. These experiences occur in us. There is, therefore, something in us that does not occur in zombies. Even if we are not *always* (and not *necessarily*) aware of these experiences, maybe we are, in introspection, sometimes aware of them. Maybe, that is, we are sometimes aware of our own awarenesses (of external things) and thus aware of the very thing (or one of the very things) that distinguishes us from zombies.[10] If we are not only aware of rocks (that is, crocks), but, sometimes at least, aware of our own awareness (experience) of rocks (the property that makes a rock a crock) then we would be aware of something (our experience of rocks) that distinguishes us from a zombie and (unlike a crock) is easily distinguished from the object (the rock) we are aware of at the perceptual level. If this is so, then one way of knowing we are not zombies is by such higher-level, introspective, awareness of our own conscious experiences.

In one sense, a perfectly trivial sense, introspection is the answer to our question. It has to be. We know by introspection that we are not zombies, that we are aware of things around (and in) us. I say this is trivial because 'introspection' is just a convenient word to describe our way of knowing what is going on in our own mind, and anyone convinced that we know—at least sometimes—what is going on in our own mind and, therefore, that we have a mind and, therefore, that we are not zombies, must believe that introspection is the answer we are looking for. I, too, believe in introspection. That is because I know—or think I know—I have thoughts, experiences and feelings. It is because I know—or think I know—that I am not a zombie, that I am aware of things. So, if I really do know what I think I know, there must be a way I know it. That—whatever it is—is what we call introspection.

This much, I say, is indisputable. I'm certainly not disputing it. The problem I have is, once again, not whether we know we are not zombies, but how we know it, how we become aware of this fact. What objects, and what properties of these objects, are we aware of that makes us aware of this fact? Do I know that I am aware of things in something like the way I know by perception, by 'outer sense,' that there is beer in the fridge—by seeing or feeling—or, if 'seeing' and 'feeling' are the wrong words for introspective awareness, by awareness of the internal experiences themselves? The way (or one of the ways) I find out there is beer in the fridge is by seeing, becoming (visually) aware of, the beer itself. I look in the fridge. I see an object—the bottle of beer—and then, by awareness of some of its revealing properties (those that reveal it to be beer), I become aware of the fact that there is beer in the fridge. I have learned that bottles of beer (and, generally speaking, only bottles of beer) look thus-and-so. Since there is something in the fridge that looks thus-and-so, I conclude (and say that I see) that there is beer in the fridge. If I don't see the beer for myself, I become aware that there is beer in the fridge (this fact) by awareness of some other object (a telltale sign, a photograph, an eyewitness report) whose features and/or behavior indicate (depend on there being) beer in the fridge. Knowledge (awareness) of external facts always depends on (perceptual) awareness of external objects and their revealing properties.[11] Telling someone how we know a fact is telling them what objects, and what properties of these objects, we became aware of that revealed this fact to us. The way I know that there is beer in the fridge, that gas is escaping, that the wine is sweet, and that the piano needs tuning, is by seeing, hearing, smelling and tasting the objects—typically (when the knowledge is direct) the beer, the gas, the wine and the piano—that 'tell me' (by their revealing properties or behaviour) that things are the way

I take them to be. It is my experience (visual, auditory, gustatory, olfactory) of these objects that 'makes' me aware of the facts. Is this the way introspection is supposed to provide an answer to our question? Do we become aware of the fact that we are aware of things, the fact that we are not zombies, by awareness of the very objects—our conscious experiences—that constitute consciousness?

In asking how you know you are not a zombie, we have to remember that I, too, know you are not a zombie. I, too, am aware of this fact. We are both aware of the fact that you are conscious. But, according to the inner sense model of introspection now being considered, you have a different way of knowing this fact than I do. Your fact-awareness (of the fact that you are aware of things) is reached by a different method than my awareness of this same fact. Your awareness of this fact is reached by some form of inner sense, by awareness of internal (to you) objects. If introspection is meant to be an answer to our question, if it is supposed to tell us not only that you know you are conscious but how you (but not I) know it, then it must, in this way, embody a claim about the special objects you are, but I am not, aware of. If our awareness of the fact that you are conscious is not reached, in this way, by a difference in the objects we are aware of, then introspection does not represent an answer to our question. It doesn't tell us how you (but not I) know this fact. It is merely a way of repeating, under the guise of a fancy label ('introspection' or 'inner sense'), that you know it.

I do not know about others, but I did not become aware of the fact that I have conscious experiences by an awareness of the conscious experiences themselves in the way I become aware of the fact that there is beer in the fridge by seeing the beer. I have experiences of beer bottles but not experiences of beer bottle experiences.[12] I think those who suppose they are introspectively aware of their own experiences are simply confusing a fact they are aware of—the fact, namely, that they have experiences—with objects they are not aware of—the experiences they have. I can be made aware (of the fact) that I am stepping on an ant by actually seeing myself step on an ant (an event) but that is not the way I become aware that I see an ant. I don't see myself see an ant. The only sense in which I am aware of myself seeing an ant is in the sense of being aware that I see an ant, but this, the awareness of the fact that I see an ant, is not my way of finding out I see an ant. It is a restatement (using the words 'aware that. . .') of the fact that I know I see an ant. If one fails to distinguish, in this way, the facts we are aware of from the objects (events) we are aware of, one will mistakenly suppose that our undisputed (by me) awareness that we experience things is an answer to a question about how we know we experience them. One will suppose, that is, that the way we become aware that we experience things is by an introspective awareness of our experience of things. This, though, is double dipping under the cloak of an equivocation: citing an awareness that *p* as one's way of becoming aware that *p*.[13]

Bill Lycan, an articulate and devoted exponent of an inner sense model of introspection, realizes that knowledge of internal facts (that, unlike zombies, we are conscious) is quite unlike ordinary sense perception. In seeing (becoming visually aware of) beer bottles, one becomes aware of some of their revealing properties. It is our awareness of these revealing properties that 'tells us' they are beer bottles. That is how we know they are beer bottles. But coming to know, by inner sense, that we have conscious experiences, Lycan tells us (1996, p. 29; 1997, p. 761), is not at all like that. Our inner sense does not reveal qualities of the objects (the experiences) being scanned. He tells us that these (first-order) experiences (of beer bottles) do not (like beer bottles) have 'ecologically significant features' and so our introspective 'scanning' of them does not represent them as having properties. The knowledge (if any) we come to have by scanning first-order conscious experiences, the internal states that distinguish us from zombies, then, is acquired without a representation of the objects—the experiences—themselves we come to know about. The experiences themselves are, so to speak, *invisible*, to the introspective scanner. We come to have knowledge of them (that we have them) without ever being made aware of them. At least we are not made aware of them, as we are of beer bottles, as objects having properties that serve to identify them.

This sounds right to me. It is Lycan's way of acknowledging that inner sense, if it makes us aware of facts about ourselves (for example, that we have conscious experiences), does so without ever making us aware of the conscious experiences themselves. But if this is right, it is also, I hasten to point out, an admission that introspection, as so understood, is not an answer to our question about how we know we have conscious experiences.[14] It tells us, at best, that

we know we are conscious, and it provides a label ('inner sense') for how we know it, but it doesn't go beyond the label and provide, as perception (of external objects) provides, propertied objects of awareness that explain how we know it. Unlike the perception of external objects and their properties (our awareness of which reveal to us the facts we come to know about them), inner sense gives us no objects and, if we follow Lycan, no properties of these objects our awareness of which explains our knowledge (awareness) of facts about them. Unless an inner sense model of introspection specifies an object of awareness whose properties (like the properties of beer bottles) indicate the facts we come to know about, an inner sense model of introspection does not tell us how we know we have conscious experiences. It merely tells us that, somehow, we know it. This is not in dispute.

We are left, then, with our original question: how do you know you are not a zombie? Not everyone who is conscious knows they are. Not everyone who is not a zombie, knows they are not. Infants don't. Animals don't. You do. Where did you learn this? To insist that we know it despite there being no identifiable *way* we know it is not very helpful. We can't do epistemology by stamping our feet.[15] Sceptical suspicions are, I think, rightly aroused by this result. Maybe our conviction that we know, in a direct and authoritative way, that we are conscious is simply a confusion of what we are aware of with our awareness of it (see Dretske forthcoming).

ACKNOWLEDGEMENTS

Thanks to Murat Aydede, Sven Bernecker, Giiven Giizeldere, and Adam Wager for helpful comments. Special thanks to Andreas Kemmerling (1999) for showing me that I didn't know I was not a zombie in the way I (Dretske 1995) thought I knew it.

REFERENCES

Armstrong, D. *Perception and the Physical World* (London: Routledge & Kegan Paul, 1961).

_____ *Bodily Sensations* (London: Routledge & Kegan Paul, 1962).

Bermúdez, J. L. *The Paradox of Self-Consciousness* (Cambridge, MA: MIT Press, Bradford Books, 1998).

Bernecker, S., "Externalism and the Attitudinal Component of Self-Knowledge," *Noûs* 30 (1996): pp. 262–75.

Burge, T., "Individualism and Self-Knowledge," *Journal of Philosophy* 85 (1988): pp. 649–93.

_____ "Our Entitlement to Self-Knowledge," *Proceedings of the Aristotelian Society* XCVI (1996): pp. 91–116.

Chalmers, D. *The Conscious Mind* (New York: Oxford University Press, 1996).

Dretske, F. *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, Bradford Books, 1981)

_____ *Naturalizing the Mind* (Cambridge, MA: MIT Press, Bradford Books, 1995).

_____ "The Mind's Awareness of Itself," *Philosophical Studies* 95 (1999): pp. 103–24. Also in *Bewusstsein und Representation,* Frank Esken and Heinz-Dieter Heckmann, eds., (Paderborn, DE: Verlag Ferdinand Schoeningh, 1997). Reprinted in Dretske 2000.

_____ *Knowledge, Perception, and Belief: Selected Essays* (Cambridge: Cambridge University Press, 2000).

Dretske, F., "Externalism and Self-Knowledge," in *New Essays on Semantic Externalism and Self-Knowledge*, S. Nuccetelli, ed., (Cambridge, MA: MIT Press, forthcoming 2003).

Gibson, J. J. *The Ecological Approach to Visual Perception* (Boston: Houghton Mifflin Publishers, 1979).

Heil, J., "Privileged Access," *Mind* 47 (1988): pp. 238–51.

Lycan, W. *Consciousness and Experience* (Cambridge, MA: MIT Press, Bradford Books, 1996).

Lycan, W., "Consciousness as Internal Monitoring," in *The Nature of Consciousness: Philosophical Debates*, N. Block, O. Flanagan, and G. Güzeldere, eds., (Cambridge, MA: MIT Press, Bradford Books, 1997).

Pitcher, G. *A Theory of Perception* (Princeton, NJ: Princeton University Press, 1971).

Reid, Thomas (1785), *Essays on the Intellectual Powers of Man*, edited and abridged by A. D. Woozley, Charlottesville: Ibis Publishing.

Siewert, C. *The Significance of Consciousness* (Princeton, NJ: Princeton University Press. 1998).

Shoemaker, S., "Introspection and the Self," in *Midwest Studies in Philosophy*, P. French, T. Uehling, and H. Wettstein, eds., (Minneapolis, MN: University of Minnesota Press, 1986), pp. 110–20. Reprinted in Shoemaker (1996) (page references are to this).

_____ "Self-Knowledge and 'Inner Sense,'" *Philosophy and Phenomenological Research* 54 (1994): pp. 249–314. Reprinted in Shoemaker (1996, pp. 201–23).

_____ *The First Person Perspective and Other Essays* (Cambridge: Cambridge University Press, 1996).

Stubenberg, L. *Consciousness and Qualia* (Amsterdam: John Benjamins Publishing, 1998).

Tye, M. *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, Bradford Books, 1995).

Wittgenstein, L. *Tractatus Logico-Philosophicus*, D. F. Pears and B. F. McGuinness, trans., (London: Routledge & Kegan Paul, 1921, 1961).

NOTES

1. For the purposes of this paper I take zombies to be human-like creatures who are not conscious and, therefore, not conscious of anything—neither objects (cars, trees, people), properties (colours, shapes, orientations), events (an object falling off the table, a sunrise), or facts (that the cup fell from the table, that the sun is rising).

   If there are readers who doubt that such creatures are possible (and it is for *this* reason they know they are not zombies), I confess to not really caring whether zombies are possible or not. Talk of zombies is merely a stylistic tool for posing an epistemological question: how do we know we are conscious? If someone believes that zombies of the kind I describe are not possible, I'm interested in how they know this. I'm also interested in whether they think this is how they know they are conscious. I doubt it.
2. Bermúdez makes a powerful case for the claim that both the self-specifying information (in perception of external objects) and the information supplied by somatic proprioception are primitive, non-conceptual, forms of self-consciousness. He acknowledges, though, that these forms of 'self-consciousness are ways of perceiving the body (the 'embodied self'), not ways of perceiving the psychologies aspects (for example, the consciousness) of the self so embodied' (1998, p. 229).
3. Thanks to Georges Dicker for suggesting this possibility. He is not responsible for the way I have formulated it.
4. Siewert 1998, pp. 19–20, 39, 172, suggests that our awareness of things (or failure to be aware of things) is what gives one first-person warrant for believing we are (or aren't) aware of things. Chalmers 1996, pp. 196–97, agrees. Neither Siewert nor Chalmers tells us *how* a conscious experience makes one conscious that one is conscious. Chalmers says that there is something 'intrinsically' epistemic about a conscious experience (ibid., p. 196). Maybe there is (though I've heard the same said about moral qualities) but our question is a question about how such experiences make one aware that one is having them. In having the experience, there need be nothing one is aware of that depends on one's having the experience. So what is it that tells one that one has an experience?
5. Those who are not direct realists about perception, and thus do not share the assumption I made in section 1, will find hallucinations, dreams, and after-images particularly relevant since, for them, veridical perception of physical objects always involves awareness of the kind of mental intermediary of which (according to some) we are aware in hallucination. Among its other targets, then, this section is meant to address the argument of those who are not direct realists, those who think that in perception we are always aware (directly) of mental particulars—items that do not occur in zombies.
6. I am not here endorsing a higher-order theory of consciousness. The question here is not what makes a mental state (for example, a feeling of thirst) conscious, but if (as higher-order theorists maintain) its being conscious consists in one's being conscious of it, how one knows that the feeling is conscious (that is, that one is conscious of it).
7. Reid 1785, p. 20: 'When I am pained, I cannot say that the pain I feel is one thing, and that my feeling it is another thing. They are one and the same thing, and cannot be disjoined, even in imagination' (taken from Stubenberg 1998, fn. 144). Shoemaker 1986, p. 20: 'Feeling pain and being in pain are, to repeat, the same thing; and the introspective knowledge that I am in pain is at the same time the introspective knowledge that I feel pain.'
8. Hereafter, in speaking of *objects* of awareness, I mean to include (as objects) events, conditions, situations, processes, activities, and states of affairs. An object, as I here use the term, is any spatio-temporal particular—for example, a sunset, a movement, a discharge, two objects standing in a certain relation to each other, a heat wave and so on. Universal properties (colours, shapes and the like) and facts (that so-and-so has such-and-such property) are not objects.
9. They would vanish in the same sense that husbands and wives would vanish if marital relations were banished. The men and women would still be there, of course; they just wouldn't be husbands and wives.
10. Although I speak indifferently (for the moment) of an awareness of x and an experience of x, there is an important (for our purposes) difference. S's awareness of x is a relationship between S and x. It includes both S and x. It is, therefore, not wholly internal to S. So we wouldn't expect S to be aware of his own awareness of x merely by inspecting internal affairs (that is, by introspection) since the targeted object of awareness—S's awareness of x—is, in part, external to S. S can't know he is aware of x unless he knows there is an x to be aware of. S's experience of x, on the other hand, is wholly internal. An experience of x is not a relationship at all, although it is described in relational terms. If there is no x, we cannot describe this internal experience as an experience of x, but the experience may, nonetheless, be the same (in all its non-relational aspects) as an experience of x. If there is no x, then we would have to describe the experience as, perhaps, an experience *as of* x (perhaps a hallucination or dream of x). So S's experience of x (unlike S's awareness of x), is something S might become aware of by inspecting internal objects. The experience of x, therefore, is an eligible object for introspective awareness.
11. There may be cases where we become aware of some fact—that, for example, there is movement on the right or that it is getting cold out—without being made aware of any object in the ordinary sense.

Peripheral vision may make us aware of a movement on the far right without making us aware of the kind of properties (shape, colour, size, orientation, etc.) usually associated with the perception of some object. Knowing that it is getting cold outside by 'feeling the cold' may be a way of becoming aware of some fact (that it is getting cold outside) by an awareness not of some object (the air? one's own body?) but, simply, the condition—the drop in temperature—one comes to know about. One feels the cold. That is how one knows it is getting cold. As indicated in note 8, I mean to include the perception of events (a drop in temperature, a movement on the right), conditions, states, etc. as instances of object perception. They are (unlike facts) spatiotemporal particulars.

12. Sydney Shoemaker 1994, 1996, especially chapters 10–12, "The Royce Lectures," gives what I regard as a definitive critique of perceptual models of self-knowledge. I recommend his essays to anyone who finds my own treatment too skimpy. I can't improve on Shoemaker.
13. For more on the distinctions between awareness of objects, awareness of properties, and awareness of facts and their conceptual (not causal) independence, see Dretske 1999.
14. In fairness, Lycan doesn't offer his inner sense theory as a theory about how we know we have conscious experiences. It is, instead, a metaphysical theory about what makes the experiences conscious.
15. At a reading of an early draft of this paper at a conference in Bielefeld, Germany, it was suggested to me by Thomas Hofweber, Thomas Grundman and Frank Hoffman that a reliability theory of knowledge provides a fairly straightforward answer to how we know we are not zombies. If we assume that zombies do not have beliefs, our belief that we are not a zombie is a reliable indication that we are not a zombie. We would not—indeed, could not—have this belief unless it was true. So the belief that one is not a zombie qualifies as knowledge that one is not a zombie. Everyone who believes they are not a zombie knows it.

As a reliability theorist, I am convinced this is the right place to look for answers, but I'm not convinced the answer is this simple. In the first place it works only for belief. It won't work if we imagine (as many do) zombies as creatures who are devoid of qualia-laden experiences (pains, tickles, visual experiences, thirst, etc.) but who may or may not have beliefs. The belief that one has conscious experiences, unlike the belief that one has beliefs, is not self-verifying. I am, furthermore, sceptical of reliability theories that identify knowledge with reliably produced belief. Unless one has grounds for one's belief, grounds that reliably indicate (carry information about) the conditions one believes to exist (perceptual experience constitutes such grounds in the case of perceptual knowledge—see Dretske 1981), beliefs, even if reliable, do not qualify as knowledge. I do not, for instance, think one gets (for free, as it were) knowledge of *P* where *P* is a physical condition necessary for life—and, therefore, necessary for belief and, in particular, the belief that *P*. Such beliefs can't be false but they aren't, not for *that* reason, knowledge. Something more is needed.

Introspection

Alex Byrne

> The word introspection need hardly be defined—it means, of course, the looking into our own minds and reporting what we there discover.
>
> —James, *The Principles of Psychology*

> 'Introspection' is a term of art and one for which little use is found in the self-descriptions of untheoretical people.
>
> —Ryle, *The Concept of Mind*

1. Introduction

I know various contingent truths about my environment by perception. For example, by looking, I know that there is a computer before me; by hearing, I know that someone is talking in the corridor; by tasting, I know that the coffee has no sugar. I know these things because I have some built-in mechanisms specialized for detecting the state of my environment. One of these mechanisms, for instance, is presently transducing electromagnetic radiation (in a narrow band of wavelengths) coming from the computer and the desk on which it sits. How that mechanism works is a complicated story—to put it mildly—and of course much remains unknown. But we can at least produce more-or-less plausible sketches of how the mechanism can start from retinal irradiation, and go on to deliver knowledge of my surroundings. Moreover, in the sort of world we inhabit,

Excerpted from *Philosophical Topics* 33:79–104, 2005. Reprinted by permission of the publisher.

specialized detection mechanisms that are causally affected by the things they detect have no serious competition—seeing the computer by seeing an idea of the computer in the divine mind, for example, is not a feasible alternative.

In addition to these contingent truths about my environment, I also know various contingent truths about my psychology. For example, I know that I *see* a computer, that I *believe* that there is someone in the corridor, that I *prefer* coffee without sugar. How do I do know these things? Well, unless it's magic, I must have some sort of mechanism (perhaps more than one) for detecting my own mental states—something rather like my visual, auditory, and gustatory systems, although directed to my mental life. That is, I have knowledge of my mental life by a special kind of *perception*, or,

> a little more cautiously, . . . something that resembles perception. But unlike sense-perception, it is not directed towards our current environment and/or our current bodily state. It is perception of the mental. Such 'inner' perception is traditionally called introspection, or introspective awareness. (Armstrong 1981, p. 60; see also Armstrong 1968, chapter 15; Lycan 1987, chapter 6; 1996, chapter 2; Nichols and Stich 2003, pp. 160–64.)

This *inner-sense theory* sounds like enlightened common sense; as Shoemaker remarks, it 'can seem a truism' (1994, p. 223). However, it is not infrequently taken to be a crass mistake.[1]

The main point of this paper is that the proponents and opponents of the inner-sense theory should split the difference. There is a mechanism for detecting one's mental states but—as will be explained later—in an important respect it does *not* 'resemble perception.'

The positive account will come at the end. The next section notes two features of self-knowledge that any theory should explain.

2. Privileged and Peculiar Access

Self-knowledge is often contrasted with knowledge of the mental states of others in the following two ways. First, knowledge of one's mental states is *privileged* in comparison to knowledge of others' minds. Roughly; beliefs about one's mental states acquired through the usual route are more likely to amount to knowledge than beliefs about others' mental states (and, more generally, beliefs about one's environment). At any rate, knowledge of one's own mental state is more likely when the state is neither factive nor object-entailing. One may well falsely believe that the cat is indoors; hence one may well falsely believe that one *knows* that the cat is indoors or *sees* that the cat is indoors. Similarly, one may well falsely believe that one sees *the cat.* But it is harder to err in believing that one *believes* that the cat is indoors, or that it *looks to one* that the cat is indoors.

To say that we have privileged access is not to say that beliefs about one's present mental states always amount to knowledge. Such beliefs need not even be true. One can falsely believe that one is angry that one wants a beer, that one believes that one is happy, for example. More controversially, one can even falsely believe that it looks to one that something is red, or that one has a headache. Nonetheless, although error may always be a possibility in a typical situation it is easier to be right about one's (non-factive, non-object-entailing) mental states (that one believes that the cat is indoors, say) than about the mental states of another (that Fred believes that the cat is indoors), or the corresponding tract of one's environment (that the cat is indoors).[2]

Second, knowledge of one's mental states is *peculiar* in comparison to one's knowledge of others' minds. One has a special method or way of knowing that one believes that the cat is indoors, that one sees the cat, that one intends to put the cat out, and so on, which one cannot use to discover that someone *else* is in the same mental state.

Our access to others' minds is importantly similar to our access to the nonpsychological aspects of our environment: one can come to know that the cat is indoors by seeing that it is, and one can likewise come to know that the cat wants to be fed and that Fred wants the sushi deluxe. Our peculiar access to our own minds is not like this: one can come to know that one wants the sushi deluxe without observing oneself at all.

Privileged and peculiar access can come apart.[3] Behaviorists typically hold that one has access to one's own mind in the same way that one has access to others' minds—by observing behavior. Yet a behaviorist might well agree that one has privileged access to one's own mind, simply because one is typically much better positioned than others to observe one's behavior.[4]

Thus Ryle:

> The superiority of the speaker's knowledge of what he is doing over that of the listener does not

> indicate that he has Privileged Access to facts of a type inevitably inaccessible to the listener, but only that he is in a very good position to know what the listener is in a very poor position to know. The turns taken by a man's conversation do not startle or perplex his wife as much as they had surprised and puzzled his fiancée, nor do close colleagues have to explain themselves to each other as much as they have to explain themselves to their new pupils. (1949, p. 171)[5]

Conversely imagine a proponent of inner sense who holds that one's 'inner eye' is very unreliable by comparison with one's outer eyes. The psychologist Karl Lashley likened introspection to astigmatic vision, claiming that '[t]he subjective view is a partial and distorted analysis' (1923, p. 338).[6] On this account, we have peculiar but underprivileged access.

The inner-sense theory does offer a nice explanation of *peculiar* access: for obvious architectural reasons, the (presumably neural) mechanism of inner sense is only sensitive to the subject's own mental states. In exactly the same style, our faculty of proprioception explains the 'peculiar access' we have to the position of our own limbs.

Self-knowledge is a large topic. To keep things manageable, the focus—following the philosophers discussed in the next three sections—will be on knowledge of one's *beliefs*. The final section briefly widens the view.

3. Self-Knowledge as Self-Constitution

This section reinforces the initial suspicion that the inner-sense theory must be right through an examination of a notable recent alternative, presented in Moran's subtle and original *Authority and Estrangement*[7] A main theme of that book is that the problem of self-knowledge is misleadingly conceived as one of 'epistemic access (whether quasi-perceptual or not) to a special realm' (Moran 2001, p. 32). In that respect, self-knowledge is unlike mathematical knowledge, knowledge of others' minds, knowledge of the past, and so on. The problem is as much one of moral psychology as it is of epistemology: we must think of '[t]he special features of first-person access. . .in terms of the special responsibilities the person has in virtue of the mental life in question being *his own*' (p. 32).

Moran's account gives a central role to the 'transparent' nature of belief, as expressed in the following well-known passage from Evans:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me 'Do you think there is going to be a third world war?' I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' (Evans 1982, p. 225)[8]

Moran first formulates Evans's observation as a 'claim of transparency':

> With respect to the attitude of belief, the claim of transparency tells us that the first-person question 'Do I believe P?' is 'transparent' to, answered in the same way as, the outward-directed question as to the truth of P itself. (Moran 2001, p. 66)

However, as Moran notes, sometimes the question 'Do I believe P?' is *not* transparent in this way, for instance in 'various familiar therapeutic contexts' (85). So the correct formulation of Evans's observation is that one can *typically* answer the question 'Do I believe P?' simply by considering whether P is true. In Moran's terminology, an answer to the question 'Do I believe P?' typically obeys the 'Transparency Condition':

> A statement of one's belief about X is said to obey the Transparency Condition when the statement is made by consideration of the facts about X itself, and not by either an 'inward glance' or by observation of one's own behavior. (p. 101)

The qualifications about no 'inward glance,' or observation of one's behavior, can be glossed as follows. Concluding that P is true is sufficient for one justifiably to claim that one believes P: no *additional* evidence—a fortiori evidence about oneself—is required.

According to Moran, transparency shows that arriving at self-knowledge (specifically, knowledge of one's beliefs) is not accurately viewed as a process of *self-discovery,* but rather as a process of *self-constitution.* Coming to know whether one believes P is not a matter of taking a 'theoretical' or disinterested stance toward oneself, of the sort one adopts toward another person when his beliefs are the subject matter of inquiry. Rather, it is a matter of 'making up one's mind' as to the truth of P. Further, transparency explains 'the special features of first-person knowledge (roughly . . . the immediacy, authority, and special relation to rationality of ordinary self-knowledge)' (Moran 2003, p. 410).

Moran's argument from transparency to the self-constitution thesis makes use of a distinction between 'theoretical' and 'practical or deliberative' questions:

> a *theoretical* question about oneself . . . is one that is answered by a *discovery of the fact of which one was ignorant*, whereas a *practical or deliberative* question is answered by a *decision or commitment of some sort* and it is not a response to ignorance of some antecedent fact about oneself. (Moran 2001, p. 58, my emphasis)

And:

> a '*deliberative*' question about one's state of mind . . . [is] a question that is answered by *making up one's mind*, one way or the other, coming to some resolution. (Moran 2003, p. 404, my emphasis)

For example, distinguish two sorts of situations in which one might ask the question 'What will I wear?' (see Moran 2001, p. 56; cf. Anscombe 1963, section 2). First, one is preparing to get dressed for the annual philosophy department party. Second, one has just been sentenced to five years for embezzling the philosophy department funds, and has yet to be issued with standard prison clothing. In the first case, the question calls for a *decision*: one considers the sartorial pros and cons, and selects the purple tie. In the second case, the question is answered by a *discovery:* the judge announces that prisoners in Massachusetts wear orange jumpsuits.

As this example shows, the distinction is not strictly speaking one between *questions*—ignoring temporal complications, it is the *same* question both times—but rather between ways of answering questions. And, indeed, Moran later writes of answering a question in 'deliberative or theoretical spirit,' taking a 'deliberative or theoretical stance' to a question, and so forth.[9] When one answers a question 'Am I F?' in a *deliberative* spirit, one engages in practical or theoretical reasoning whose outcome (a belief in the theoretical case, an action/intention in the practical case) determines either that one is F, or that one is not F. That is, the outcome of one's reasoning determines the answer. When one answers a question 'Am I F?' in a *theoretical* spirit, one engages in theoretical reasoning whose outcome is simply to uncover the answer, not to determine it.

The distinction applies to questions like 'Do I believe P?' One might address this question in a theoretical spirit, treating it 'as a more or less purely psychological question about a certain person, as one may enquire into the beliefs of someone else' (Moran 2001, p. 67). Alternatively, one might address this question in a deliberative spirit, as a matter of making up one's mind about P. Take, for example, the question 'Do I believe Alice is a threat to my career?' as asked by her colleague Bert. After looking back over his behavior toward Alice—anonymously rejecting one of Alice's papers that criticizes Bert's pet theory, etc., etc.—Bert might conclude that he has this belief. Alternatively, Bert might address the question in a deliberative spirit, and investigate whether Alice really is a threat to Bert's career. Perhaps the result of the investigation is that Alice is harmless, and Bert thereby concludes that he believes that Alice is not a threat. We can imagine Bert addressing the question in both a deliberative and theoretical spirit, raising the uncomfortable possibility of discovering that he has inconsistent beliefs.[10]

Here is how Moran links the 'deliberative/theoretical' distinction with transparency:

> With respect to belief, the claim of transparency is that from within the first-person perspective, I treat the question of my belief about P as equivalent to the question of the truth of P. What I think we can see now is that the basis for this equivalence hinges on the role of deliberative considerations about one's attitudes. For what the 'logical' claim of transparency requires is the deferral of the theoretical question 'What do I believe?' to the deliberative question 'What am I to believe?' And in the case of the attitude of belief, answering a deliberative question is a matter of determining what is true.
>
> When we unpack the idea in this way, we see that the vehicle of transparency in each case lies in the requirement that I address myself to the question of my state of mind in a *deliberative* spirit, deciding and declaring myself on the matter, and not confront the question as a purely psychological one about the beliefs of someone who happens also to be me. (p. 63)

Suppose I ask myself 'Do I believe P?' and that I answer 'I believe P' by determining that P is true. Then, according to Moran, I have answered this question by 'a decision or commitment of some sort,' and not 'by a discovery of the fact of which I was ignorant.' Transparency shows, in other words, that knowledge that one believes P, when arrived at by considering whether P is true, is a matter of 'making up one's mind' that P is true.

However, Moran's conclusion is overdrawn. It is true that often one answers the question

'Do I believe P?' in a deliberative spirit. It is natural to imagine this happening with Evans's question 'Do you think there is going to be a third world war?' One has not previously considered the likelihood of a third world war; one studies the relevant geopolitical facts, and makes up one's mind. But, precisely because it suggests this sort of context, in this respect Evans's example is misleading. Consider the question 'Do I believe that I live in Cambridge, Massachusetts?' or 'Do I believe that Moran is the author of *Authority and Estrangement*?' These questions can be answered transparently, by considering the relevant facts of location and authorship, but I do not need to make up my mind.[11] On the contrary, it is already made up. I have believed for some time that I live in Cambridge, and that Moran is the author of *Authority and Estrangement* I can know that I believe I live in Cambridge, for example, by remembering the nonpsychological fact that I live in Cambridge.

So transparency does *not* show that knowledge of one's beliefs is in general a matter of making up one's mind. And there are further reasons to be suspicious of any tight connection between transparency and the thesis that self-knowledge involves self-constitution. Moran concentrates almost exclusively on the transparency of belief, but perception provides other examples (as noted in Evans 1982, pp. 224–25; see also Dretske 1995, 2003). One can know that one sees the cat by an 'outward look' at the cat. One determines that the cat is there, and concludes that one sees the cat. However, *seeing the cat* is not in *any* sense a matter of making up one's mind, or 'coming to some resolution'—one can see the cat without having any beliefs about it. When one comes to know that one sees the cat by looking at the cat, one has simply discovered 'some antecedent fact about oneself.'

Although the significance of Moran's special cases should not be overlooked, self-knowledge of mental states in general (or even of beliefs in particular), with its distinctive features of privileged and peculiar access, cannot possibly be explained in terms of 'self-constitution.' Moran's emphasis on *transparency* is quite another matter, however, and much will be made of that later.[12]

. . . But there is another way in which self-knowledge might involve reasoning without the perception of anything mental. Recall the quotation from Evans about belief and transparency (section 3). One apparently finds out that one believes that it's raining by determining that it's raining: knowledge that one has this belief, insofar as it rests on perceptual evidence at all, rests on perceptual evidence about the weather, not on perceptual evidence of one's behavior or anything mental. That is, one *reasons* from the evidence that it's raining, to the conclusion that one believes that it's raining. If this procedure can yield self-knowledge, and if it involves the (causal) *detection* of the belief that it's raining, then this would be an instance of the 'broad perceptual model' without being either Ryleanism or the inner-sense theory.

And although we haven't yet found a compelling reason to reject the inner-sense theory, there is a reason for pursuing an alternative. As noted in section 2, the inner-sense theory neatly explains *peculiar* access. But it does not explain *privileged* access. In fact, it leaves it something of a mystery. Why is inner sense less prone to error than the outer senses? (Recall Lashley's astigmatic inner eye.) And why is there not (or, at any rate, not obviously) an actual psychological condition that approximates Shoemaker's 'self-blindness'?

7. Transparent Rules

The account to follow appeals to the notion of *following a rule*, specifically an *epistemic* rule. This apparatus of epistemic rules needs to be explained first.

7.1. Epistemic Rules

Holmes's reasoning to the conclusion that Mr. White killed Mr. Orange is complex, and his methods resist easy summary. Presumably Holmes's reasoning is somehow rule-governed, but it is not clear how to identify the rules. On the other hand, some reasoning is considerably simpler. For example, Mrs. Hudson might hear the doorbell ring, and conclude that there is someone at the door. By hearing that the doorbell is ringing, Mrs. Hudson knows that the doorbell is ringing; by reasoning, she knows that there is someone at the door.

It is natural to say that Mrs. Hudson acquires knowledge of her visitors by following a simple recipe or rule. If we say that an *epistemic rule* is a conditional of the following form:

R If conditions C obtain, believe that p[13]

then the epistemic rule that Mrs. Hudson follows is:

> DOORBELL If the doorbell rings, believe that there is someone at the door[14]

What does it mean to say that Mrs. Hudson *follows* this rule on a particular occasion? For present purposes this semi-stipulative answer will suffice: Mrs. Hudson believes that there is someone at the door *because* she recognizes that the doorbell is ringing. The 'because' is intended to mark the kind of reason-giving causal connection that is often discussed under the rubric of 'the basing relation'. Mrs. Hudson might recognize that the doorbell is ringing, and believe that there is someone at the door for some *other* reason; in this case, she does not form her belief because she recognizes that the doorbell is ringing.

So S follows the rule R ('If conditions C obtain, believe that p') on a particular occasion if on that occasion:

(i) S believes that p because she recognizes that conditions C obtain which implies:
(ii) S recognizes (hence knows) that conditions C obtain
(iii) conditions C obtain
(iv) S believes that p

Following DOORBELL tends to produce knowledge about one's visitors (or so we may suppose), and hence it is a *good* rule. Following *bad* rules tends to produce false and unjustified beliefs, for example:

> NEWS If the *Weekly World News* reports that p, believe that p

NEWS is also an example of a *schematic* rule. One *follows* a schematic rule just in case one follows a rule that is an instance of the schematic rule; a schematic rule is *good* to the extent that its instances are.

If the antecedent conditions C of an epistemic rule R are not specified in terms of the rule follower's mental states, R is *neutral* A schematic rule is neutral just in case some of its instances are. Thus, the claim that S can follow a neutral rule does not presuppose that S has the capacity for self-knowledge. DOORBELL and NEWS are neutral rules; 'If you intend to go swimming, believe that you will get wet' is not.[15]

Self-knowledge is our topic, not skepticism: knowledge of one's environment (including others' actions and mental states) and reasoning (specifically, rule-following of the kind just sketched) can be taken for granted. So, in the present context, it is not in dispute that we follow neutral rules, including neutral rules with mentalistic fillings for 'p,' like 'If S has a rash, believe that S feels itchy'; neither is it in dispute that some neutral rules are good rules.

Moran's 'claim of transparency' (section 3) can be recast using the apparatus of epistemic rules as follows. Knowledge of one's beliefs may be obtained by following the neutral schematic rule:

> BEL If p, believe that you believe that p

Since the antecedent of BEL expresses the content of the mental state that the rule-follower ends up believing she is in, BEL can be called a *transparent* rule.

7.2. The Puzzle of Transparency

But how can following BEL lead to self-knowledge? In his contribution to a symposium on *Authority and Estrangement*, Moran acutely observes that there is a puzzle here:

> the claim of Transparency *is* something of a paradox: how can a question referring to a matter of empirical psychological feet about a particular person be legitimately answered without appeal to the evidence about that person, but rather by appeal to a quite independent body of evidence? (2003, p. 413)

This *puzzle of transparency* can be expressed in the terminology of epistemic rules as follows. Apparently, knowledge of what one believes is often the result of following the neutral schematic rule BEL, yet surely this is a *bad* rule: that p is the case does not even make it *likely* that one believes that it is the case.[16]

However, recall the 'rule of necessitation' in modal logic. According this rule of inference, if a sentence 'p' is a line of a proof, one may write down the necessitation of 'p,' '$\Box p$,' as a subsequent line. Artificially forcing this into a format similar to that of 'epistemic rules,' the rule of necessitation becomes:

> NEC if 'p' is a line, you may write '$\Box p$' as a subsequent line

NEC, it seems, does not preserve truth, and so—in an extended sense—is a 'bad' rule. It doesn't follow from the fact that the cat is indoors that *necessarily* the cat is indoors. The cat's being indoors doesn't even make it *likely* that this state of affairs could not have been otherwise.

But, of course, the rule of necessitation is not a bad rule. In fact, it's a necessarily

truth-preserving rule. The reason is that—assuming that the only initial premises of a proof are axioms—whenever one is in a position to follow the rule by writing down '$\Box p$' 'p' *is* a necessary truth. The axioms of a system of modal logic are themselves necessary truths, and whatever follows from them by the other rules are also necessary truths. So whenever one is in circumstances in which the rule applies—whenever, that is, one is confronted with a proof whose initial premises are axioms—every line of the proof is a necessary truth. If the allowable substituends for 'p' include sentences about the location of cats, then the rule of necessitation is a bad rule. But if (as intended) it is kept within the confines of modal logic, the rule is perfectly good.

Having noticed this, it is a short step to noticing that something analogous holds for BEL. One is only in a position to follow BEL by believing that one believes that p when one has recognized that p. And recognizing that p is (inter alia) coming to *believe* that p.[17] BEL is *self-verifying* in this sense: if it is followed, the resulting second-order belief is true. Compare a third-person version of BEL:

BEL-3 If p, believe that Fred believes that p

BEL-3 is of course not self-verifying: the result of following it maybe (indeed, is very likely to be) a false belief about Fred's beliefs.

Given that we follow rules like DOORBELL, it should not be in dispute that we *can* follow BEL. Given the plausibility of Evans's observation about the procedure we actually follow, it should not be in dispute that we *do* follow BEL. The puzzle of transparency is solved by noting that BEL is self-verifying; since the goodness of rules like DOORBELL can be assumed, it should not be in dispute that following BEL will often produce *knowledge* of what one believes. BEL offers an obvious explanation of *peculiar* access: as just noted, BEL-3 is a very bad rule indeed. But, if BEL is to be the whole story, privileged access must also be explained. At a minimum, we need to show that BEL is significantly *better*—more knowledge-conducive—than rules whose consequents concern others' mental states.

7.3. Privileged Access Explained

Since following a rule like DOORBELL will deliver beliefs that are about as likely to amount to knowledge as our beliefs about others' mental states, for simplicity DOORBELL can go proxy for a good rule whose consequent concerns others' mental states. In what ways is BEL better than DOORBELL?

One immediate advantage of BEL over DOORBELL is that the former but not the latter is self-verifying. Suppose one follows DOORBELL, and so knows that the doorbell is ringing and believes that there is someone at the door. One's belief that there is someone at the door is probably true, but it may be false. Suppose one also follows BEL: in particular, one recognizes that the doorbell is ringing and thereby believes that one believes that the doorbell is ringing. Because BEL is self-verifying, the truth of one's second-order belief is guaranteed.

Suppose there is someone at the door, and so the belief produced by following DOORBELL is true—how likely is it to be knowledge? Following Sosa (1999) and Williamson (2001, chapter 5), say that one's belief that p is *safe* just in case one's belief could not easily have been false.[18] Safety is a plausible necessary condition for knowledge; absent countervailing considerations (such as having excellent but misleading evidence that not-p), safety can be used as a rough-and-ready diagnostic tool for the presence of knowledge where the proposition in question is contingent. Could one easily have been wrong about the presence of a visitor? The ways in which one could have falsely believed that there is someone at the door can be classified into three types:

> *Type I:* not-p, and one falsely believes that conditions C obtain, thereby believing that p. Perhaps the sound made by a passing ice cream truck might have been mistaken for the ringing of the doorbell, leading to the false belief that there is someone at the door.
>
> *Type II*: not-p, and one truly believes that conditions C obtain, thereby believing that. Perhaps a wiring defect might have caused the doorbell to ring, leading to the false belief that there is someone at the door.
>
> *Type III*: not-p, and one believes that p, but not because one knows or believes that conditions C obtain. Perhaps too much coffee might have lead one to believe that there is someone at the door, even if the stoop had been deserted.

By hypothesis, there is someone at the door. Also by hypothesis, one *follows* DOORBELL, which entails one *knows* that the doorbell is ringing. Hence one could not easily have been wrong about that, and so Type I errors are remote possibilities. And, given certain assumptions that will obtain in many realistic cases (the doorbell has no wiring defects, the coffee is not

that psychoactive, etc.), Type II and Type III errors are also remote possibilities and could not easily have happened. However, in other realistic cases these errors *are* nearby possibilities, and hence one's true belief that there is someone at the door will not be knowledge.

Consider now the belief that one believes that there is someone at the door; could one easily have been wrong? It is not possible to make a Type I error: one cannot falsely believe that the doorbell is ringing without believing that the doorbell is ringing. Type II errors are likewise ruled out: one cannot truly believe that the doorbell is ringing without believing that the doorbell is ringing.

If one follows BEL, only Type III errors are a threat to one's knowledge: perhaps too much coffee would have lead one to believe that one believes that the doorbell is ringing, even if one had not believed that the doorbell is ringing. With the modest assumption that Type III errors are equally likely when following BEL as when following DOORBELL, the true beliefs produced by following BEL are more likely to amount to knowledge than the true beliefs produced by following DOORBELL.[19]

Sometimes one will not succeed in following DOORBELL because one believes but does not know that the doorbell is ringing (maybe a passing ice cream truck induces a false belief). Say that S *tries* to follow rule R if S believes that *p* because S *believes* that conditions C obtain. That S follows R entails that she tries to follow R, but not conversely. If one tries to follow DOORBELL but does not succeed, then one will not *know* that there is someone at the door; if one's belief about a visitor is true, that is just an accident. The visitor could have easily been delayed, with the truck passing as it actually did, in which case one would have falsely believed that there is someone at the door. That is, a Type I error is a nearby possibility.

Sometimes one will not succeed in following BEL either: one will merely try to follow it, and believe but not know that the doorbell is ringing. But one's second order belief that one believes that the doorbell is ringing will be *true*. As before, Type I and II errors are not possible. Hence this situation will be commonplace: trying to follow BEL, one investigates whether *p, mistakenly* concludes that *p*, and thereby comes to *know* that one believes that *p*. (In these cases, one will know that one believes that *p* on the basis of no evidence at all.)

BEL, then, has considerable epistemic virtues, but it is important to not overstate them. Consider the following quotation from Evans (continuing the quotation given in section 3 above):

> I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*. . . If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states: even the most determined sceptic cannot find here a gap in which to insert his knife. (Evans 1982, p. 225)

This passage can be read as suggesting that (a) the ability to investigate whether *p* brings with it the ability to find out whether one believes that *p* (assuming one has the concept of belief), and that (b) following BEL cannot fail to produce knowledge of one's beliefs.

Whether or not this is Evans's view, on the present account it is incorrect.[20] There is no guarantee that one can follow (or try to follow) BEL, no matter how talented one is at gaining knowledge of cats, doorbells, other minds, and so on. Neither is there any guarantee that the beliefs produced by following (or trying to follow) BEL will amount to knowledge.

However, as a contingent matter, trying to follow BEL will usually produce knowledge of what one believes. Venturing out on a limb—of course the matter requires more discussion—we may tentatively conclude that privileged access is thereby explained.

7.4. Shoemaker's Insights Validated

The hypothesis that privileged and peculiar access is explained by our following (or trying to follow) BEL *is* a version of Shoemaker's 'broad perceptual model.' Suppose someone follows an instance of BEL, and thereby believes that she believes that it's raining. She looks out of the window, say, and sees that it's raining. A causal transition between mental states occurs, as it does when one believes that there is someone at the door because one has recognized that the doorbell is ringing. The subject believes (knows) that it's raining, which causes her to believe that she believes that it's raining. Thus, there is an appropriate causal mechanism. Further, the state detected is independent of its detection. The subject might not have followed BEL, in which case the first-order belief would have been present without the second-order belief. What's more (we may fairly suppose), someone might believe that it's raining, possess the concept of

belief, and yet not even have the capacity to follow BEL.

Yet the account is not a version of the inner-sense theory. It is *economical*, like behaviorism. Taking the capacity to follow good neutral rules for granted, knowledge of what one believes comes along more-or-less for free. Since this capacity belongs to the department of reasoning, not perceiving, Shoemaker's idea that the source of self-knowledge can be traced to 'rationality' is vindicated, albeit not via his preferred route.

8. Beyond Belief?

It is often pointed out the phenomenon of transparency is quite limited.[21] It covers belief and perception, but what about, to take the obvious example, knowledge of one's *wants* or *desires*? I come to know that I *believe* that I have a beer by looking outward and discovering a beer in my hand. One does not typically come to know that one *wants* a beer by the same procedure. Typically, when I want a beer, that is because I *don't* have one. And I often have privileged and peculiar access, when in such a beerless condition, to the fact that I want a beer. So some other account is required to explain one's knowledge of one's desires, to say nothing of wishes, hopes, fears, expectations, thoughts, intentions, imaginings, and the rest.[22]

So far, none of this is an *objection* to the account of the previous section, just some observations that highlight the limitations of the approach. But there is an impending problem—the *puzzle of opacity*. Suppose that the epistemology of nontransparent states is *extravagant*, and hence that self-knowledge of wants, hopes, intentions, and so forth cannot be explained in terms of our ability to follow neutral rules. In short: a faculty of introspection is needed. Then the puzzle is this: why isn't inner sense ever operative in the case of transparent mental states? Why is the epistemology of these states always (apparently) of the transparent and economical sort?

Perhaps the puzzle of opacity can be solved, consistently with an extravagant epistemology of nontransparent mental states. But a two-tiered account of self-knowledge—economical in the case of transparent mental states, extravagant in the case of the rest—is not an entirely comfortable position. This at least motivates an exploration of whether extravagance is really needed.

Let us briefly consider desire. There are desires and desires: likes, wants, preferences, cravings, lusts, wishes, and so on. A first-person epistemology for all members of this heterogeneous category will not be attempted here. Instead, as a tractable example, take the preference for one of a range of options. I have neither beer nor wine, and am offered one or the other. Here is the beer, the culmination of centuries of Belgian brewing tradition. There is the wine, the product of my host's home winemaking hobby. I prefer the beer to the wine. How do I know that?

As just noted, clearly transparency does not apply here—I do not (usually) find out that I prefer the beer by finding out that I am holding a glass of the stuff. However, often my eyes are still 'directed outward—upon the world.' I can investigate my preferences by attending to the *beer* and the *wine*, and their relative merits (and perhaps to the host as well, in particular her tendency to take offense). I conclude that the beer wins over the wine, and thereby conclude that I prefer the beer. The relevant (neutral) rule is roughly this:

> DES If ψing is a better option than χing, believe that you prefer to ψ than to χ[23]

DES is a neutral rule, so the capacity to follow it does not presuppose the capacity for self-knowledge. But is DES a *good* rule? One's preferences tend to line up with one's beliefs about the merits of the options: if one believes that ψing is a better option than χing, one typically prefers to ψ than to χ. On the face of it, DES is a good rule; moreover, it can be used to explain privileged access to one's desires in a way that closely (although not perfectly) parallels the explanation given in section 7.3 above.

Unlike BEL, DES is not a transparent rule. And there is another important difference. No doubt *Socrates* would be happy with the claim that one prefers to ψ than to χ whenever one believes that ψing is a better option than χing, but—as is familiar from the literature on weakness of will—he appears to have been wrong about that. Unlike BEL, DES is not self-verifying. Slavishly following (or trying to follow) DES will sometimes lead to a *false* belief about one's mental state.

Consider an example. Suppose I believe that the wine is a better option than the beer, because my host will take offense if I choose the beer. Nonetheless, I selfishly prefer the beer to the wine. If I follow DES, I will falsely believe that my preferences are the other way round. The

problem is not that such a mistake can never happen—it can. (I believe that it would be considerably better to read *Mind and World* this evening than to watch *The Real World;* thinking myself a person who is sensitive to intellectual virtues, I believe that I prefer the former to the latter. However, I find myself turning on the television, leaving the book unopened.)[24] The problem, rather, is that sometimes one *knows* that one's preferences are at odds with one's better judgments. In particular, despite believing that the wine is better than the beer (or, alternatively, not having an opinion on the matter), I may well have privileged and peculiar access to the fact that, all things considered, I prefer the beer to the wine.

The issue is whether this sort of knowledge will require an extravagant faculty of introspection. And here the buck is passed from preference to intention, because in the situation just described, it is plausible that I know I prefer the beer to the wine because I know that I *intend* to have the beer.

Further investigation will have to be deferred.[25] To summarize the conclusion so far: at least with respect to belief, the inner-sense theory is partly right. There is an inner mechanism for detecting one's beliefs. But the inner-sense theory is also partly wrong: the mechanism comes with our capacity for reasoning about the external world—there is no inner eye.

REFERENCES

Allan, L. G., "Time Perception," in *Encyclopedia of Psychology*, A. Kazdin, ed., (Washington, DC: American Psychological Association, 2000).

Alston, W., "Varieties of Privileged Access," *American Philosophical Quarterly* 8 (1971): pp. 223–41.

Anscombe, G. E. M. *Intention* (Oxford: Blackwell, 1963), 2nd ed.

Armstrong, D. M. *A Materialist Theory of the Mind* (London: Routledge & Kegan Paul, 1968).

Armstrong, D. M. *The Nature of Mind and Other Essays* (Ithaca, NY: Cornell University Press, 1981).

Arpaly, N. *Unprincipled Virtue* (Oxford: Oxford University Press, 2003).

Aydede, M., "Is Introspection Inferential?" in *Privileged Access: Philosophical Accounts of Self-Knowledge*, B. Gertler, ed., (Aldershot: Ashgate Publishing, 2003).

Bar-On, D. *Speaking My Mind* (Oxford: Oxford University Press, 2005).

Bern, D. J., "Self-Perception Theory," in *Advances in Experimental Social Psychology*, vol. 6, L. Berkowitz, ed., (New York: Academic Press, 1972).

Boghossian, P., "Content and Self-Knowledge," *Philosophical Topics* 17 (1989): pp. 5–26. Page reference to the reprinting in Ludlow and Martin 1998.

Byrne, A. *Transparency and Self-Knowledge* (Oxford: Oxford University Press, 2020).

Dretske, F., "Introspection," *Proceedings of the Aristotelian Society* 94 (1994): pp. 263–78.

_____ *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995).

_____ "How Do You Know You Are Not a Zombie?" in *Privileged Access: Philosophical Accounts of Self-Knowledge*, B. Gertler, ed., (Aldershot: Ashgate Publishing, 2003).

Evans, G. *The Varieties of Reference* (Oxford: Oxford University Press, 1982).

Finkelstein, D. H. *Expression and the Inner* (Cambridge, MA: Harvard University Press. 2003).

Gertler, B., "Self-Knowledge," in *The Stanford Encyclopedia of Philosophy*, E. Zalta, ed., (Spring 2003), http://plato.stanford.edu/archives/spr2003/entries/self-knowledge/.

Goldman, A., "The Psychology of Folk Psychology," *Behavioral and Brain Sciences* 16 (1993): pp. 15–28.

_____ "The Mentalizing Folk," in *Metarepresentations*, D. Sperber, ed., (Oxford: Oxford University Press, 2000).

_____ "Internalism Exposed," in *Knowledge, Truth, and Duty*, M. Steup, ed., (Oxford: Oxford University Press, 2001).

Gordon, R., "'Radical' Simulationism," in *Theories of Theories of Mind*, P. Carruthers and P. Smith, eds., (Cambridge: Cambridge University Press, 1996).

Lashley, K., "The Behavioristic Interpretation of Consciousness II," *Psychological Review* 30 (1923): pp. 329–53.

Ludlow, P., and Martin, N., eds., *Externalism and Self-Knowledge* (Stanford, CA: CSLI, 1998).

Lycan, W. G. *Consciousness* (Cambridge, MA: MIT Press, 1987).

_____ *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996).

_____ "Dretske's Ways of Introspecting," in *Privileged Access: Philosophical Accounts of Self-knowledge*, B. Gertler, ed., (Aldershot: Ashgate Publishing, 2003).

Lyons, W. *The Disappearance of Introspection* (Cambridge, MA; MIT Press, 1986).

Martin, M., "An Eye Directed Outward," in *Knowing Our Own Minds*, C. Wright, B. Smith, and C. Macdonald, eds., (Oxford: Oxford University Press, 1998).

Moran, R. *Authority and Estrangement* (Princeton, NJ: Princeton University Press, 2001).

_____ "Responses to O'Brien and Shoemaker," *European Journal of Philosophy* 11 (2003): pp. 402–19.

Nichols, S., and Stich, S. *Mindreading* (Oxford: Oxford University Press, 2003).

Peacocke, C. *A Study of Concepts* (Cambridge, MA: MIT Press, 1992).

_____"Conscious Attitudes, Attention, and Self-Knowledge," in *Knowing Our Own Minds*, C. Wright, B. Smith, and C. Macdonald, eds., (Oxford: Oxford University Press, 1998).

Ryle, G. *The Concept of Mind* (London: Hutchinson, 1949). Page reference to the 1980 Penguin reprint.

Shoemaker, S., "On Knowing One's Own Mind," *Philosophical Perspectives* 2 (1988): pp. 183–209. Page reference to the reprinting in Shoemaker 1996.

_____"Self-Knowledge and 'Inner-Sense.'" *Philosophy and Phenomenological Research* 54 (1994): pp. 249–314. Page reference to the reprinting in Shoemaker 1996.

_____ *The First-Person Perspective and Other Essays* (Cambridge: Cambridge University Press, 1996).

Sosa, E. "How to Defeat Opposition to Moore," *Philosophical Perspectives* 13 (1999): pp. 141–53.

Thomasson, A., "Introspection and Phenomenological Method," *Phenomenology and the Cognitive Sciences* 2 (2003): pp. 239–54.

Williamson, T. *Knowledge and Its Limits* (Oxford: Oxford University Press, 2001).

Wilson, T. *Strangers to Ourselves: Discovering the Adaptive Unconscious* (Cambridge, MA: Harvard University Press, 2002).

Wright, C., "Self-Knowledge: The Wittgensteinian Legacy," in *Knowing Our Own Minds*, C. Wright, B. Smith, and C. Macdonald, ed., (Oxford: Oxford University Press, 1998).

NOTES

Distant ancestors of parts of this paper were included in talks given at Alberta, Calgary, Stanford, USC, the University of Texas at Austin, Union College, Western Washington University, Vermont, and a Metaphysics and Epistemology conference in Dubrovnik, Croatia. The numerous comments I received on those occasions greatly improved this paper. I am especially grateful to the participants in Ned Block and Thomas Nagel's seminar on language and mind at New York University, and in my graduate seminars on other minds and self-knowledge at MIT. For comments on the penultimate draft, thanks to David Hilbert, Richard Holton, Ed Minar, and Susanna Siegel.

1. Cf. Wright 1998, p. 24: 'The privileged observation explanation [of 'first-third person asymmetries in ordinary psychological discourse'] is unquestionably a neat one. What it *does* need philosophy to teach is its utter hopelessness.'
2. As much psychological research has shown, we are often mistaken about our reasons for belief or action (Wilson 2002); however, none of this undermines the (relatively modest) sort of privileged access claimed in the text. For some discussion of what this research does and doesn't show, see Wilson 2002, pp. 104–15; see also Nichols and Stich 2003, p. 161.
3. Indeed, they actually come apart for *knowing that the cat is indoors, seeing the cat,* and the like; we have peculiar but not (an impressive kind of) privileged access to these states.
4. The claim of 'Privileged Access,' in Ryle's sense, is this: '(1). . . a mind cannot help being constantly aware of all the supposed occupants of its private stage, and (2). . . it can also deliberately scrutinize by a species of non-sensuous perception at least some of its own states and operations' (1949, p. 148) In the contemporary literature, 'privileged access' is often used approximately for what (in the text) is described as privileged *and* peculiar access (see, e.g., Alston 1971; Moran 2001, pp. 9–10).
5. For the near-Rylean position in psychology, see Bern 1972. Although one (perhaps incautious) statement of Ryle's official view is that 'in principle, as distinct from practice, John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe' (1949, p. 149), he has no account of the third-personal method by which (to take Ryle's own examples) 'I can catch myself daydreaming,' or 'catch myself engaged in a piece of silent soliloquy' (p. 160).
6. This quotation, together with Lashley's comparison with astigmatism, appears in Lyons 1986, p. 29. Ryle's characterization of 'inner perception' denies that there are 'any counterparts to deafness, astigmatism' (1949, p. 157).
7. Another equally notable recent alternative is Bar-On 2005, to be passed over for reasons of space.
8. See also Dretske 1994, 1995; and Gordon 1996. A similar view can be found in Flusserl: see Thomasson 2003 for an interesting discussion.
9. See Moran 2001, pp. 63, 64, 65, 67.
10. At one point, Moran contrasts the two ways of answering the question 'Do I believe P?' as follows:

    > In characterizing the two sorts of questions one may direct towards one's state of mind, the term 'deliberative' is best seen at this point in contrast to 'theoretical,' the primary point being to mark the difference between that enquiry which terminates in a true description of my state, and one which terminates in the formulation or endorsement of an attitude. (2001, p. 63)

    However, this is misleading (and is not Moran's considered view). In successfully answering the question 'Do I believe P?,' whether in a deliberative or theoretical spirit, one comes to have a true belief about one's beliefs, and so in both cases the enquiry 'terminates in a true description of [one's] state.'
11. For what is essentially the same point, see Peacocke 1998, pp. 215–16.
12. As will become apparent, the account given in section 7 classifies Moran's special cases together with examples where one's mind is already made up, as both involving (in Moran's phrase) 'epistemic access. . . to a special realm.'
13. Since judging is the act that results in the state of belief, perhaps the consequent is better put as 'judge that *p*.' This is simply a stylistic or presentational issue, however. The linguistic formulation of the rule only plays a heuristic role—*all* the work is done by the account of *following* a rule (see immediately below).

14. No doubt the epistemological story is considerably more complicated; DOORBELL should be treated as a harmless simplification.
15. 'You' refers to the rule-follower; tenses are to be interpreted so that the time the rule is followed counts as the present.
16. The *locus classicus* for the puzzle of transparency (as it arises for perception) is Dretske 2003; see also Martin 1998, pp. 117–18.
17. Of course, there are many differences between the rule of necessitation and BEL. For one thing, logical rules of inference are not rules of reasoning. With this cautionary remark in mind, there is another point of analogy. It is a mistake to think that the rule of necessitation is equivalent to the (invalid) axiom schema '$p \supset \Box p$' plus modus ponens; likewise, it is a mistake to think of following BEL as equivalent to (falsely) assuming that for all *P*, if *P* is true then one believes *P*, which would make one's reasoning from the premise that it's raining to the conclusion that one believes that it's raining demonstrative.

    In his earlier writings on transparency, Dretske likened the phenomenon to examples of 'displaced perception' such as the following: I see that the bathroom scale on which I am standing reads '170' and infer that I weigh 170 pounds (Dretske 1994, p. 263; 1995, p. 41). This reasoning stands or falls with the prior reasonableness of the assumption (or 'connecting belief' [1995, 42]) that if the bathroom scale on which I am standing reads '170' then I weigh 170 pounds, and so is importantly *dis*analogous to BEL. Against Dretske, Aydede 2003 complains, in effect, that the connecting beliefs in the mental case are often false (see also Lycan 2003, pp. 16–17, and 26–27, n. 1). (For Dretske's account of the difference between the mental case and examples like the scale, see Dretske 1995, pp. 60–61.)
18. The formulation in the text is (approximately) Williamson's. For simplicity, situations that could easily have obtained in which one does not falsely believe that *p* but rather falsely believes something else will be ignored. See Williamson 2001, pp. 101–2.
19. Some with 'internalist' sympathies might insist that considerations of safety and the like are not enough: if following BEL leads to knowledge, the knowledge-conducive properties of BEL have to be in some way 'accessible' to the rule-follower. This issue is just an instance of the general debate between externalism and internalism (see, e.g., Goldman 2001). On the face of it, the present proposal does not make an internalist account of self-knowledge *especially* problematic (compared to, say, an internalist account of perceptual knowledge); accordingly the externalism/internalism debate need not be examined here.
20. In the case of perception, which he contrasts with belief, Evans does deny that transparency 'produce[s] infallible knowledge' (1982, p. 228).
21. See Goldman 2000, pp. 182–83; Nichols and Stich 2003, p. 194; Finkelstein 2003, postscript; Bar-On 2004, pp. 114–18.
22. The BEL-style approach cannot be applied to perception without modification. For example, the rule 'If *p*, believe that you see ('visually') that *p*' is not good. One often knows that the cat is indoors without seeing that it is.
23. Something like this view is implicit in Moran 2001 (see Finkelstein 2003, p. 161). A similar suggestion is also made in Gertler 2003, section 2.3. DES and some of the subsequent points in the text can be extracted from Shoemaker 1988, p. 47–48.
24. For a probing discussion of other sorts of examples, see Arpaly 2003.
25. See Byrne, 2020.

21 The Unreliability of Naive Introspection

Eric Schwitzgebel

i.

Current conscious experience is generally the last refuge of the skeptic against uncertainty. Though we might doubt the existence of other minds, that the sun will rise tomorrow, that the earth existed five minutes ago, that there's any 'external world' at all, even whether two and three make five, still we can know, it's said, the basic features of our ongoing stream of experience. Descartes espouses this view in his first two *Meditations*. So does Hume, in the first book of the *Treatise*, and—as I read him—Sextus Empiricus.[1] Other radical skeptics like Zhuangzi and Montaigne, though they appear to aim at very general skeptical goals, don't grapple specifically and directly with the possibility of radical mistakes about current conscious experience. Is this an unmentioned exception to their skepticism? Unintentional oversight? Do they dodge the issue for fear that it is too poor a field on which to fight their battles?[2] Where is the skeptic who says: We have no reliable means of learning about our own ongoing

From *Philosophical Review* 117: 245–273.

conscious experience, our current imagery, our inward sensations—we are as in the dark about that as about anything else, perhaps even more in the dark?

Is introspection (if that's what's going on here) just that good? If so, that would be great news for the blossoming—or should I say recently resurrected?—field of consciousness studies. Or does contemporary discord about consciousness—not just about the *physical bases* of consciousness but seemingly about the basic features of experience itself—point to some deeper, maybe fundamental, elusiveness that somehow escaped the notice of the skeptics, that perhaps partly explains the first, ignoble death of consciousness studies a century ago?

ii.

One must go surprisingly far afield to find major thinkers who unambiguously hold, as I do, that the introspection of current conscious experience is both (i) possible, important, necessary for a full life, and central to the development of a full scientific understanding of the mind, and (ii) highly untrustworthy. In Eastern meditative traditions, I think this is a commonplace. Also the fiercest advocates of introspective training in the first era of scientific psychology and 'phenomenology' (circa 1900) endorsed both claims—especially E. B. Titchener.[3] Both the meditators and Titchener, though, take comfort in optimism about introspection 'properly' conducted—so they hardly qualify as *general* skeptics or pessimists. It's as though their advocacy of a regimen sets them free to criticize introspection as ordinarily practiced. Might they be right in their doubts but less so in their hopes? Might we need introspection, though the prospects are bleak?

I won't say much to defend (i), which I take to be both common sense and the majority view in philosophy. Of course we have *some* sort of attunement to our ongoing conscious experience, and we impoverish ourselves if we try to do without it. Part (ii) is the project. In less abbreviated form: Most people are poor introspectors of their own ongoing conscious experience. We fail not just in assessing the *causes* of our mental states or the processes underwriting them; and not just in our judgments about nonphenomenal mental states like traits, motives, and skills; and not only when we are distracted, or passionate, or inattentive, or self-deceived, or pathologically deluded, or when we're reflecting about minor matters, or about the past, or only for a moment, or where fine discrimination is required. We are both ignorant and prone to error. There are major lacunae in our self-knowledge that are not easily filled in, and we make gross, enduring mistakes about even the most basic features of our currently ongoing conscious experience (or 'phenomenology'), even in favorable circumstances of careful reflection, with distressing regularity. We either err or stand perplexed, depending—rather superficially, I suspect—on our mood and caution. (This essay will focus on error, but sufficient restraint can always transform error to mere ignorance.)

Contemporary philosophers and psychologists often doubt the layperson's talent in assessing such nonconscious mental states as personality traits, motivations and skills, hidden beliefs and desires, and the bases of decisions; and they may construe such doubts as doubts about 'introspection.' But it's one thing not to know why you chose a particular pair of socks (to use an example from Nisbett and Wilson 1977) and quite another to be unable accurately to determine your currently ongoing *visual experience* as you look at those socks, your *auditory experience* as the interviewer asks you the question, the experience of pain in your back making you want to sit down. Few philosophers or psychologists express plain and general pessimism about the latter sorts of judgment. Or, rather, I should say this: I have heard such pessimism only from *behaviorists*, and their near cousins, who nest their arguments in a theoretical perspective that rejects the psychological value, sometimes even the coherence, of attempting to introspect conscious experiences at all—and thus reject claim (i) above—though indeed even radical behaviorists often pull their punches when it comes to ascribing flat-out *error*.[4]

Accordingly, though infallibilism—the view that we cannot err in our judgments about our own current conscious experience—is now largely out of favor, mainstream philosophical criticism of it is meek. Postulated mistakes are largely only momentary, or about matters of fine detail, or under conditions of stress or pathology, or at the hands of malevolent neurosurgeons.[5] Fallibilists generally continue to assume that, in favorable circumstances, careful introspection can reliably reveal at least the broad outlines of one's currently ongoing experience. Even philosophers most of the community sees

as radical are, by my lights, remarkably tame and generous when it comes to assessing our accuracy in introspecting current conscious experience. Paul Churchland 1985, 1988 puts it on a par with the accuracy of sense perception. Daniel Dennett 2002 says that we can come close to infallibility when we are charitably interpreted.[6] Where are the firebrands?

A word about 'introspection.' I happen to regard it as a species of attention to currently ongoing conscious experience, but I won't defend that view here. The project at hand stands or falls quite independently. Think of introspection as you will—as long as it is the primary method by which we normally reach judgments about our experience in cases of the sort I'll describe.[7] That method, whatever it is, is unreliable as typically executed. Or so I will argue in this essay.

iii.

I don't know what emotion is, exactly. Neither do you, I'd guess. Is surprise an emotion? Comfort? Irritability? Is it more of a gut thing, or a cognitive thing? Assuming cognition isn't totally irrelevant, how is it involved? Does cognition relate to emotion merely as cause and effect, or is it somehow, partly, constitutive?

I'm not sure there's a single right answer to these questions. The empirical facts seem ambiguous and tangled.[8] Probably we need to conjecture and stipulate, simplify, idealize, to have anything workable. So also, probably, for most interesting psychological concepts. But here's one thing that's clear: Whatever emotion is, some emotions—joy, anger, fear—can involve or accompany conscious experience.

Now, you're a philosopher, or a psychologist, presumably interested in introspection and consciousness and the like, or you wouldn't be reading this article. You've had emotional experiences, and you've thought about them, reflected on how they feel as they've been ongoing or in the cooling moments as they fade. If such experiences are introspectible, and if introspection is the diamond clockwork often supposed, then you have some insight. So tell me: Are emotional states like joy, anger, and fear always felt phenomenally—that is, as part of one's stream of conscious experience—or only sometimes? Is their phenomenology, their experiential character, always more or less the same, or does it differ widely from case to case? For example, is joy sometimes in the head, sometimes more visceral, sometimes a thrill, sometimes an expansiveness—or, instead, does joy have a single, consistent core, a distinctive, identifiable, unique experiential character? Is emotional consciousness simply the experience of one's bodily arousal, and other bodily states, as William James 1981 (1890) seems to suggest? Or, as most people think, can it include, or even be exhausted by, something less literally visceral? Is emotional experience consistently located in space (for example, particular places in the interior of one's head and body)? Can it have color—for instance, do we sometimes literally 'see red' as part of being angry? Does it typically come and pass in a few moments (as Buddhists sometimes suggest), or does it tend to last awhile (as my English-speaking friends more commonly say)?

If you're like me, you won't find all such questions trivially easy. You'll agree that someone—perhaps even yourself—could be mistaken about some of them, despite sincerely attempting to answer them, despite a history of introspection, despite maybe years of psychotherapy or meditation or self-reflection. You can't answer these questions one-two-three with the same easy confidence that you can answer similarly basic structural questions about cars—how many wheels? hitched to horses? travel on water? If you can—well heck, I won't try to prove you wrong! But if my past inquiries are indicative, you are in a distinct minority.

It's not just *language* that fails us—most of us?—when we confront such questions (and if it were, we'd have to ask, anyway, why this particular linguistic deficiency?) but introspection itself. The questions challenge us not simply because we struggle for the *words* that best attach to a patently obvious phenomenology. It's not like perfectly well knowing what particular shade of tangerine your Volvo is, stumped only about how to *describe* it. No, in the case of emotion the *very phenomenology itself*—the 'qualitative' character of our consciousness—is not entirely evident, or so it seems to me. But how could this be so, if we know the 'inner world' of our own experience so much better than the world outside? Even the grossest features of emotional experience largely elude us. Reflection doesn't remove our ignorance, or it delivers haphazard results.

Relatedly, most of us have a pretty poor sense, I suspect, of what brings us pleasure and suffering. Do you really enjoy Christmas? Do you really feel bad while doing the dishes? Are

you happier weeding or going to a restaurant with your family? Few people make a serious study of this aspect of their lives, despite the lip service we generally pay to the importance of 'happiness.' Most people feel bad a substantial proportion of the time, it seems to me.[9] We are remarkably poor stewards of our emotional experience. We may *say* we're happy—overwhelmingly we do—but we have little idea what we're talking about.[10]

iv.

Still, you might suggest, when we attend to *particular instances* of ongoing emotional experience, we can't go wrong, or don't, or not by far. We may concede the past to the skeptic, but not the present. It's impossible—nearly impossible?—to imagine my being wrong about my ongoing conscious experience *right now*, as I diligently reflect.

Well, philosophers *say* this, but I confess to wondering whether they've really thought it through, contemplated a variety of examples, challenged themselves. You'd hope they would have, so maybe I'm misunderstanding or going wrong in some way here. But to me at least, on reflection, the claim that I could be infallible in everything I'm inclined to say about my ongoing consciousness—even barring purely linguistic errors, and even assuming I'm being diligent and cautious and restricting myself to simple, purely phenomenal claims, arrived at (as far as I can tell) 'introspectively'—well, unfortunately that just seems blatantly unrealistic.

Let's try an experiment. You're the subject. Reflect on, introspect, your own ongoing emotional experience at this instant. Do you even have any? If you're in doubt, vividly recall some event that still riles you until you're sure enough that you're suffering some renewed emotion. Or maybe your boredom, anxiety, irritation, or whatever in reading this essay is enough. Now let me ask: Is it completely obvious to you what the character of that experience is? Does introspection reveal it to you as clearly as visual observation reveals the presence of the text before your eyes? Can you discern its gross and fine features through introspection as easily and confidently as you can, through vision, discern the gross and fine features of nearby external objects? Can you trace its spatiality (or nonspatiality), its viscerality or cognitiveness, its involvement with conscious imagery, thought, proprioception, or whatever, as sharply and infallibly as you can discern the shape, texture, and color of your desk? (Or the difference between 3 and 27?) I cannot, of course, force a particular answer to these questions. I can only invite you to share my intuitive sense of uncertainty. (Perhaps I can buttress this sense of uncertainty by noting, in passing, the broad range of disputes and divergences within the literature on the experiential character of emotion—disputes that at least *seem* to be about emotional phenomenology itself, not merely about its causes and connections to nonexperiential states, or about how best to capture it in a theory.[11])

Or consider this: My wife mentions that I seem to be angry about being stuck with the dishes again (despite the fact that doing the dishes makes me happy?). I deny it. I reflect; I sincerely attempt to discover whether I'm angry—I don't just reflexively defend myself but try to be the good self-psychologist my wife would like me to be—and still I don't see it. I don't think I'm angry. But I'm wrong, of course, as I usually am in such situations: My wife reads my face better than I introspect. Maybe I'm not quite *boiling* inside, but there's plenty of angry phenomenology to be discovered if I knew better how to look. Or do you think that every time we're wrong about our emotions, those emotions must be nonconscious, dispositional, not genuinely *felt*? Or felt and perfectly apprehended phenomenologically but somehow nonetheless mislabeled? Can't I also err more directly?

Surely my 'no anger' judgment is colored by a particular self-conception and emotional involvement. To that extent, it's less than ideal as a test of my claim that, even in the most favorable circumstances of quiet reflection, we are prone to err about our experience. However, as long as we focus on judgments about emotional phenomenology, such distortive factors will probably be in play. If that's enough consistently to undermine the reliability of our judgments, that rather better supports my thesis than defeats it, I think.

Infallible judges of our emotional experience? I'm baffled. How could anyone believe that? Do *you* believe that? What am I missing?

v.

Now maybe emotional experience is an unusually difficult case. Maybe, though we err there, we are generally quite accurate in our

judgments about *other* aspects of our phenomenology. Maybe my argument even plays on some conceptual confusion about the relation between emotion and its phenomenology or relies illegitimately on introspection's undercutting the emotion introspected. I don't think so, but I confess I have no tidy account to eradicate such worries.

So let's try vision. Suppose I'm looking directly at a nearby, bright red object in good light, and I judge that I'm having the visual phenomenology, the 'inward experience,' of redness. Here, perhaps—even if not in the emotional case—it seems rather hard to imagine that I could be wrong in that judgment (though I could be wrong in using the term 'red' to *label* an experience I otherwise perfectly well know).

I'll grant that. Some aspects of visual experience are so obvious it would be difficult to go wrong about them. So also would it be difficult to go wrong in some of our judgments about the external world—the presence of the text before your eyes, the existence of the chair in which you're sitting and are now (let's suppose) minutely examining. Introspection may admit obvious cases, but that in no way proves that it's more secure than external perception—or even as secure.

Now of course many philosophers have argued plausibly that one *could* be wrong even in 'obvious' judgments about external objects, if one allows that one may be dreaming or that one's brain may have been removed at night and teleported to Alpha Centauri to be stimulated by genius neuroscientists with inputs mimicking normal interaction with the world. Generally, philosophers have supposed (with Descartes) that such thought experiments don't undermine judgments about visual phenomenology. So perhaps obvious introspective judgments *are* more secure than obvious perceptual ones, after all, since they don't admit even this peculiar smidgen—usually it only seems like a smidgen—of doubt?

But in dreams we make baldly incoherent judgments, or at least very stupid ones. I think I can protrude my tongue without its coming out; I think I see red carpet that's not red; I see a seal as my sister without noticing any difficulty about that. In dream delirium, these judgments may seem quite ordinary or even insightful. If you admit the possibility that you're dreaming, I think you should admit the possibility that your judgment that you are having reddish phenomenology is a piece of delirium, unaccompanied by any actual reddish phenomenology. Indeed, it seems to me not entirely preposterous to suppose that we have no color experiences at all in our sleep—or have them only rarely—and our judgments about the colors of dream objects are on par with the seal-sister judgment, purely creative fiction unsupported by any distinctive phenomenology.[12] If so, the corresponding judgments about the coloration of our *experiences* of those dream objects will be equally unsupported.

Likewise, if malevolent neurosurgeons from Alpha Centauri may massage and stoke our brains, I see no reason to deny them the power to produce directly the judgment that one is having reddish phenomenology, while suppressing the reddish phenomenology itself. Is this so patently impossible?[13]

Absolute security and immunity to skeptical doubt thus elude even 'obvious' introspective judgments as well as perceptual ones. If we rule out radically skeptical worries, then we're left with judgments on a par ('red phenomenology now,' 'paper in my hands')—judgments as obvious and secure as one could reasonably wish. The issue of whether the introspection of current visual experience warrants greater trust than the perception of nearby objects must be decided on different grounds.

vi.

Look around a bit. Consider your visual experience as you do this. Does it seem to have a center and a periphery, differing somehow in clarity, precision of shape and color, richness of detail? Yes? It seems that way to me, too. Now consider this: How *broad* is that field of clarity? Thirty degrees? More? Maybe you're looking at your desk, as I am. Does it seem that a fairly wide swath of the desk—a square foot?—presents itself to you clearly in experience at any one moment, with the shapes, colors, textures all sharply defined? Most people endorse something like this view when I ask them.[14] They are, I think, mistaken.

Consider, first, our visual capacities. It's firmly established that the precision with which we detect shape and color declines precipitously outside a central, foveal area of about one to two degrees of arc (about the size of your thumbnail held at arm's length). Dennett 1991 has suggested a way of demonstrating this to yourself. Draw a card from a normal deck without looking at it. Keeping your eyes fixed on some point in front of you, hold the

card at arm's length just beyond your field of view. Without moving your eyes, slowly rotate the card toward the center of your visual field. How close to the center must you bring it before you can determine the color of the card, its suit, and its value? Most people are quite surprised at the result of this little experiment. They substantially overestimate their visual acuity outside the central, foveal region. When they can't make out whether it's a Jack or a Queen, though the card is nearly (but only *nearly*) dead center, they laugh, they're astounded, dismayed.[15] You have to bring it *really* close.

By itself, this says nothing about our visual experience. Surprise and dismay may reveal error in our normal (implicit) assumptions about our visual capacities, but it's one thing to mistake one's abilities and quite another to misconstrue phenomenology. Our visual experience depends on the recent past, on general knowledge, on what we hear, think, and infer, as well as on immediate visual input—or so it's plausible to suppose. Background knowledge could thus fill in and sharpen our experience beyond the narrow foveal center. Holding our eyes still and inducing ignorance could artificially crimp the region of clarity.

Still, I doubt visual experience is nearly as sharp and detailed as most untutored introspectors seem to think. Here's the root of the mistake, I suspect: When the thought occurs to you to reflect on some part of your visual phenomenology, you normally *move your eyes* (or 'foveate') in that direction. Consequently, wherever you think to attend, within a certain range of natural foveal movement, you find the clarity and precision of foveal vision. It's as though you look at your desk and ask yourself: Is the stapler clear? Yes. The pen? Yes. The artificial wood grain between them and the mouse pad? Yes—each time looking directly at the object in question—and then you conclude that they're all clear simultaneously.[16]

But you *needn't* reflect in this way. We can prize foveation apart from introspective attention. Fixate on some point in the distance, holding your eyes steady while you reflect on your visual experience outside the narrow fovea. Better, direct your introspective energies away from the fovea while your eyes continue to move around (or 'saccade') normally. This may require a bit of practice. You might start by keeping one part of your visual field steadily in mind, allowing your eyes to foveate anywhere but there. Take a book in your hands and let your eyes saccade around its cover, while you think about your visual experience in the regions away from the precise points of fixation.

Most of the people I've spoken to, who attempt these exercises, eventually conclude to their surprise that their experience of clarity decreases substantially even a few degrees from center. Through more careful and thoughtful introspection, they seem to discover—in fact, I think they really do discover—that visual experience does not consist of a broad, stable field, flush with precise detail, hazy only at the borders. They discover that, instead, the center of clarity is tiny, shifting rapidly around a rather indistinct background. My interlocutors—most of them—confess to error in having originally thought otherwise.

If I'm right about this, then most naive introspectors are badly mistaken about their visual phenomenology when they first reflect on it, when they aren't warned and coached against a certain sort of error, even though they may be patiently considering that experience as it occurs. And the error they make is not a subtle one; the two conceptions of visual experience differ vastly. If naive introspectors are as wrong as many later confess to be about the clarity and stability of visual experience, they're wrong about an *absolutely fundamental and pervasive* aspect of their sensory consciousness.

I'm a pretty skeptical guy, though. I'm perfectly willing to doubt myself. Maybe I'm wrong and visual experience is a plenum. But if so, I'm not the only person who's wrong about this. So also are most of my interlocutors (whom I *hope* I haven't browbeaten too badly) and probably a good number of philosophers and psychologists.[17] We—I, my friends and cobelievers—have been seduced into error by some theory or preconception, perhaps, some blindness, stupidity, oversight, suggestibility. Okay, let's assume that. I need only, now, turn my argument on its head. We *tried* to get it right. We reflected, sincerely, conscientiously, in good faith, at a leisurely pace, in calm circumstances, without external compulsion, and we got it wrong. Introspection failed us. Since what I'm trying to show is the aptitude of introspection to lead to just such errors, that result would only further my ultimate thesis. Like other skeptical arguments that turn on our capacity for disagreement, it can triumph in partial defeat.

I do have to hold this, though: Our disagreement is real and substantial. My interlocutors' opinions about their ongoing visual experience change significantly as a result of their

reflections. The mistake in question, whichever side it's on, though perhaps understandable, is large—no miniscule, evanescent detail, no mere subtlety of language. Furthermore, opinions on both sides arise from normal introspective processes—the same types of process (whatever they are) that underwrite most of our 'introspective' claims about consciousness. And finally, I must hold that those who disagree don't differ in the basic structure of their visual experience in such a way as to mirror precisely their disagreements. Maybe you can successfully attack one of these premises?

vii.

In 2002, David Chalmers and David Hoy ran a summer seminar in Santa Cruz, California, for professional philosophers of mind. They dedicated an entire week of the seminar to the 'phenomenology of intentionality,' including most centrally the question of whether *thought* has a distinctive experiential character.

There can be little doubt that sometimes when we think, reflect, ruminate, dwell, or what have you, we simultaneously, or nearly so, experience *imagery* of some sort: maybe visual imagery, such as of keys on the kitchen table; maybe auditory imagery, such as silently saying 'that's where they are.' Now here's the question to consider: Does the phenomenology of thinking *consist entirely* of imagery experiences of this sort, perhaps accompanied by feelings (emotions?) such as discomfort, familiarity, confidence? Or does it go beyond such images and feelings? Is there some distinctive phenomenology specifically of thought, additional to, or conjoined with, the images, perhaps even capable of transpiring without them?

Scholars disagree. Research and reflection generate dissent, not convergence, on this point. This is true historically,[18] and it was also true at the Santa Cruz seminar: Polled at the week's end, seventeen participants endorsed the existence of a distinctive phenomenology of thought, while eight disagreed, either disavowing the phenomenology of thought altogether or saying that imagery exhausts it.[19]

If the issue were highly abstract and theoretical, like most philosophy, or if it hung on recondite empirical facts, we might expect such disagreement. But the introspection of current conscious experience—that's supposed to be *easy*, right? Thoughts occupied us throughout the week, presumably available to be discerned at any moment, as central to our lives as the seminar table. If introspection can guide us in such matters—if it can guide us, say, at least as reliably as vision—shouldn't we reach agreement about the existence or absence of a phenomenology of thought as easily and straightforwardly as we reach agreement about the existence of the table?

Unless people diverge so enormously that some have a phenomenology of thought and others do not, then someone is quite profoundly mistaken about her own stream of experience. Disagreement here is no matter of fine nuance. If there is such a thing as a conscious thought, then presumably we have them all the time. How could you go looking for them and simply not find them? Conversely, if there's no distinctive phenomenology of thought, how could you introspect and come to believe that there is—that is, invent a whole category of conscious experiences that simply don't exist? Such fundamental mistakes almost beggar the imagination; they plead for reinterpretation as disagreements only in language or theory, not real disagreements about the phenomenology itself.

I don't think that's how the participants in these disputes see it, though; and, for me at least, the temptation to recast it this way dissipates when I attempt the introspection myself. Think of the Prince of Wales. Now consider: Was there something it was like to have that thought? Set aside any visual or auditory imagery you may have had. The question is: Was there something further in your experience, something besides the imagery, something that might qualify as a distinctive phenomenology of thinking? Try it again, if you like. Is the answer *so obvious* you can't imagine someone going wrong about it? Is it as obvious as that your desk has drawers, your shirt is yellow, your shutters are cracked? Must disagreements about such matters *necessarily* be merely linguistic or about philosophical abstracta? Or, as I think, might people genuinely misjudge even this very basic, absolutely fundamental and pervasive aspect of their conscious experience, even after putting their best introspective resources to work?

viii.

In my view, then, we're prone to gross error, even in favorable circumstances of extended reflection, about our ongoing emotional, visual,

and cognitive phenomenology. Elsewhere, I've argued for a similar ineptitude in our ordinary judgments about auditory experience and visual imagery. I won't repeat those arguments here.[20] All this is evidence enough, I think, for a generalization: The introspection of current conscious experience, far from being secure, nearly infallible, is faulty, untrustworthy, and misleading—not just *possibly* mistaken, but massively and pervasively. I don't think it's just me in the dark here, but most of us. You too, probably. If you stop and introspect now, there's likely very little you should confidently say you know about your own current phenomenology. Perhaps the right kind of learning, practice, or care could largely shield us from error—an interesting possibility that merits exploration!—but I see as yet no robust scientific support for such hopes.[21]

What about pain, a favorite example for optimists about introspection? Could we be infallible, or at least largely dependable, in reporting ongoing pain experiences? Well, there's a reason optimists like the example of pain—pain and foveal visual experience of a single bright color. It *is* hard, seemingly, to go too badly wrong in introspecting really vivid, canonical pains and foveal colors. But to use *these cases only* as one's inference base rigs the game. And the case of pain is not always as clear as sometimes supposed. There's confusion between mild pains and itches or tingles. There's the football player who sincerely denies he's hurt. There's the difficulty we sometimes feel in locating pains precisely or in describing their character. I see no reason to dismiss, out of hand, the possibility of genuine introspective error in these cases. Psychosomatic pain, too: Normally, we think of psychosomatic pains as genuine *pains*, but is it possible that some, instead, involve sincere belief in a pain that doesn't actually exist?

Inner speech—'auditory imagery' as I called it above—can also seem hard to doubt—that I'm silently saying to myself 'time for lunch.' But on closer inspection, I find it slipping from my grasp. I lean toward thinking that there is a conscious phenomenology of imageless thought (as described in sec. 7)—but as a result, I'm not always sure whether some cogitation that seems to be in inner speech is not, instead, imageless. And also: Does inner speech typically involve not just auditory images but also motor images in the vocal apparatus? Is there an experiential distinction between inner speaking and inner hearing? I almost despair.

Why, then, do people tend to be so confident in their introspective judgments, especially when queried in a casual and trusting way? Here's my suspicion: Because no one ever scolds us for getting it wrong about our experience and we never see decisive evidence of error, we become cavalier. This lack of corrective feedback encourages a hypertrophy of confidence. Who doesn't enjoy being the sole expert in the room whose word has unchallengeable weight? In such situations, we tend to take up the mantle of authority, exude a blustery confidence—and genuinely *feel* that confidence (what professor doesn't know this feeling?) until we imagine possibly being proven wrong later by another authority or by unfolding events. About our own stream of experience, however, there appears to be no such humbling danger.

ix.

But wait. Suppose I say, 'I'm thinking of a pink elephant'—or even, simply, 'I'm thinking.' I'm sincere, and there's no linguistic mistake. Aren't claims of this sort necessarily self-verifying? Doesn't merely thinking such thoughts or reaching such judgments, aloud or silently, guarantee their truth? Aren't, actually, their *truth* conditions just a subset of their *existence* conditions?—and if so, mightn't this help us out somehow in making a case for the trustworthiness of introspection?

I'll grant this: Certain things plausibly follow from the very having of a thought: that I'm thinking, that I exist, that something exists, that my thought has the content it in fact has. Thus, certain thoughts and judgments will be infallibly true whenever they occur—whatever thoughts and judgments assert the actuality of the conditions or consequences of one's having them. But the general accuracy of introspective judgments doesn't follow.

Infallibility is, in fact, cheap. *Anything* that's evaluable as true or false, if it asserts the conditions or consequences of its own existence or has the right self-referential structure, can be infallibly true. The spoken assertion 'I'm speaking' or 'I'm saying 'blu-bob'' is infallibly true whenever it occurs. The sentence 'This sentence has five words' is infallibly true whenever uttered. So is the semaphore assertion 'I'm holding two flags.' So, sure, certain thoughts are infallibly true—true whenever they occur. This shouldn't surprise us; it's merely an instance of the more general phenomenon of self-fulfillment. It has

nothing whatsoever to do with introspection; it implies no perfection in the art of ascertaining what's going on in one's mind. If introspection happens to be the process by which thoughts of this sort sometimes arise, that's merely incidental: Infallibly self-fulfilling thoughts are automatically true whether they arise from introspection, from fallacious reasoning, from evil neurosurgery, quantum accident, stroke, indigestion, divine intervention, or sheer frolicsome confabulation.

And how many introspective judgments, really, *are* infallibly self-fulfilling? 'I'm thinking'—okay. 'I'm thinking of a pink elephant'—well, maybe, if we're liberal about what qualifies as 'thinking of' something.[22] But 'I'm not angry,' 'my emotional phenomenology right now is entirely bodily,' 'I have a detailed image of the Taj Mahal, in which every arch and spire is simultaneously well defined,' 'my visual experience is all clear and stable one hundred degrees into the periphery,' 'I'm having an imageless thought of a pink elephant.' Those are a different matter entirely, I'd say.

And, anyway, I'm not so sure we haven't changed the topic. Does the thought 'I'm thinking' or 'I'm thinking of a pink elephant' really express a judgment about *current conscious experience*? Philosophers might reasonably take different stands here, but it's not clear to me that I'm committed to believing *anything*, or anything particular, about my conscious experience in accepting such a judgment. I'm certainly not committed to thinking I have a visual image of a pink elephant, or an 'imageless thought' of one, or that the words 'pink elephant' are drifting through my mind in inner speech. I might hold 'I'm thinking of a pink elephant' to be true while I suspect any or all of the latter to be false. Am I committed at least to the view that I'm *conscious*? Maybe. Maybe this is one fact about our conscious experience we infallibly know. (Could I reach the judgment that I'm conscious *nonconsciously*?)[23] But your ambitions for introspection must be modest indeed if that satisfies you.

X.

I sometimes hear the following objection: When we make claims about our phenomenology, we're making claims about how things *appear* to us, not about how anything actually *is*. The claims, thus divorced from reality, can't be false; and if they're true, they're true in a peculiar way that shields them from error. In looking at an illusion, for example, I may well be wrong if I say the top line is longer; but if I say it *appears* or *seems* to me that the top line is longer, I can't in the same way be wrong. The sincerity of the latter claim seemingly guarantees its truth. It's tempting, perhaps, to say this: If something *appears* to appear a certain way, necessarily it appears that way. Therefore, we can't misjudge appearances, which is to say, phenomenology.

This reasoning rests on an equivocation between what we might call an *epistemic* and a *phenomenal* sense of 'appears' (or, alternatively, 'seems'). Sometimes, we use the phrase 'it appears to me that such-and-such' simply to express a judgment—a hedged judgment, of a sort—with no phenomenological implications whatsoever. If I say, 'It appears to me that the Democrats are headed for defeat,' ordinarily I'm merely expressing my opinion about the Democrats' prospects. I'm not attributing to myself any particular phenomenology. I'm not claiming to have an image, say, of defeated Democrats, or to hear the word 'defeat' ringing in my head. In contrast, if I'm looking at an illusion in a vision science textbook, and I say that the top line 'appears' longer, I'm *not* expressing any sort of judgment about the line. I know perfectly well it's not longer. I'm making instead, it seems, a claim about my phenomenology, about my visual experience.[24]

Epistemic uses of 'appears' *might* under certain circumstances be infallible in the sense of the previous section. Maybe, if we assume that they're sincere and normally caused, their truth conditions will be a subset of their existence conditions—though a story needs to be told here.[25] But *phenomenal* uses of 'appears' are by no means similarly infallible. This is evident from the case of weak, nonobvious, or merely purported illusions. Confronted with a perfect cross and told there may be a 'horizontal-vertical illusion' in the lengths of the lines, one can feel uncertainty, change one's mind, and make what at least plausibly seem to be errors about whether one line 'looks' or 'appears' or 'seems' in one's visual phenomenology to be longer than another. You might, for example, fail to notice—or worry that you may be failing to notice—a real illusion in your experience of the relative lengths of the lines; or you might (perhaps under the influence of a theory) erroneously report a minor illusion that actually isn't part of your visual experience at all. Why not?[26]

Philosophers who speak of 'appearances' or 'seemings' in discussing consciousness invite conflation of the epistemic and phenomenal senses of these terms. They thus risk breathing an illegitimate air of indefeasibility into our reflections about phenomenology. 'It appears that it appears that such-and-such' may have the look of redundancy, but on disambiguation the redundancy vanishes: 'it epistemically seems to me that my phenomenology is such-and-such.' No easy argument renders this statement self-verifying.

xi.

Suppose I'm right about one thing—about something that appears, anyway, hard to deny: that people reach vastly different introspective judgments about their conscious experience, their emotional experience, their imagery, their visual experience, their thought. If these judgments are all largely *correct*, people must differ immensely in the structure of their conscious experience.

You might be happy to accept that if the price of denying it is skepticism about introspective judgments. Yet I think there's good reason to pause. Human variability, though impressive, usually keeps to certain limits. Feet, for example—some are lean and bony, some fat and square, yet all show a common design: skin on the outside, stout bones at the heel, long bones running through the middle into toes, nerves and tendons arranged appropriately. Only in severe injury or mutation is it otherwise. Human livers may be larger or smaller, better or worse, but none is made of rubber or attached to the elbow. Human behavior is wonderfully various, yet we wager our lives daily on the predictability of drivers, and no one shows up to department meetings naked. Should phenomenology prove the exception by varying radically from person to person—some of us experiencing one hundred degrees of visual clarity, some only two degrees, some possessed of a distinctive phenomenology of thought, some lacking it, and so forth—with as little commonality as these diverse self-attributions seem to suggest? Of course, if ocular physiology differed in ways corresponding to the differences in report, or if we found vastly different performances on tests of visual acuity or visual memory, or if some of us possessed higher cognition or sympathetic emotional arousal while others did not, that would be a different matter. But as things are, two people walk into a room, their behavioral differences are subtle, their physiologies are essentially the same, and yet phenomenologically they're so alien as to be like different species? Hmm!

Here's another possibility: Maybe people are largely the same *except* when they introspect. Maybe we all have basically the same visual phenomenology most of the time, for example, until we reflect directly on that phenomenology—and then some of us experience one hundred degrees of stable clarity while others experience only two degrees. Maybe we all have a phenomenology of thought, but introspection amplifies it in some people, dissipates it in others; analogously for imagery, emotions, and so forth.

That view has its attractions. But to work it so as to render our introspective judgments basically trustworthy, one must surrender many things. The view concedes to the skeptic that we know little about ordinary, unintrospected experience since it hobbles the inference from introspected experience to experience in the normal, unreflective mode. It threatens to make a hash of change in introspective opinion: If someone thinks a previous introspective opinion of hers was mistaken—a fairly common experience among people I interview (see, for example, section vi)—she must, it seems, generally be *wrong* that it was mistaken. She must, generally, be correct, now, that her experience is one way, and also correct, a few minutes ago, that it was quite another way, without having noticed the intervening change. This seems an awkward coupling of current introspective acumen with profound ignorance of change over time. The view renders foolish whatever uncertainty we may sometimes feel when confronted with what might have seemed to be introspectively difficult tasks (as in sections iv, vii, and x). Why feel uncertain if the judgment one reaches is bound to be right? It also suggests a number of particular—and I might say rather doubtful—empirical commitments (unless consciousness is purely epiphenomenal): major differences in actual visual acuity while introspecting between those reporting broad clarity and those reporting otherwise; major differences in cognition while introspecting between people reporting a phenomenology of thought and those denying it; and so on. The view also requires an entirely different explanation of why theorists purporting to use 'immediate retrospection'[27] also find vastly divergent results—since immediate retrospection, if successful, postpones the act of

introspection until after the conscious experience to be reported, when presumably it won't have been polluted by the introspective act.

Is there some compelling reason to take on all this?

xii.

There are two kinds of unreliability. Something might be unreliable because it often goes wrong or yields the wrong result, or it might be unreliable because it fails to do anything or yield any result at all. A secretary is unreliable in one way if he fouls the job, unreliable in another if he neglects it entirely. A program for delivering stock prices is unreliable in one way if it tends to misquote, unreliable in another if it crashes. Either way, they can't be depended on to do what they ought.[28]

Introspection is unreliable in both ways. Reflection on basic features of ongoing experience leads sometimes to error and sometimes to perplexity or indecision. Which predominates in the examples of this essay is not, I think, a deep matter, but rather a matter of context or temperament. Some introspectors will be more prone to glib guesswork than others. Some contexts—for example, a pessimistic essay on introspection—will encourage restraint. But whether the result is error or indecision, introspection will have failed—*if* we suppose that introspection ought to yield trustworthy judgments about the grossest contours of ongoing conscious experience.

You might reject that last idea. Maybe we shouldn't expect introspection to reveal (for example) the bodily or nonbodily aspects of emotion, the presence or absence of a distinctive cognitive phenomenology. It wouldn't, then, tell against the reliability of introspection if such cases baffle us. It doesn't tell against the reliability of a stock quote program if it doesn't describe the weather. A passenger car that overheats going 120 mph isn't thereby unreliable. Maybe I've pushed introspection beyond its proper limits, illegitimately forcing it into failure.

What, then, would be the proper domain of introspection, construed narrowly enough to preserve its reliability? Our ongoing beliefs and desires? That changes the topic away from current conscious experience: When I report believing that a bodybuilder is governor of California, I'm not, I think—at least not directly and primarily—reporting introspectively on an ongoing episode of consciousness.[29] Our current thoughts and emotions—but only their contents, not their form or structure? That, too, might be changing the topic. Thought and emotion may not be best construed as purely phenomenal. The self-attribution of current thought contents and emotional states (as opposed to the phenomenal form and structure of those thoughts and emotions) may be more expressive or reactive (like a spontaneous 'I hate you!') or simply self-fulfilling (section ix) than introspective if we're going to be strict about what properly falls in the domain of introspection. And of course the accuracy of emotional self-attribution is disputable (sec. 4), as, I think, is the accuracy of our self-attribution of recently past thought contents.[30]

We may generally be right about foveal visual experience of color and the presence or absence of canonical pains, but it's arbitrary to call such reports introspective and not similar-seeming reports about the overall clarity of the visual field or the presence or absence of bodily aspects of emotion. In both formal and informal interviews with me, and in the experiments of early introspective psychologists like Titchener 1901–5) and in the recent explorations of psychologists like Hurlburt 1990; Hurlburt and Schwitzgebel 2007, subjects confidently pronounce on the features of experience discussed in this essay. Neither I, nor they, nor Titchener, nor Hurlburt, nor anyone else I'm aware of, sees any obvious difference in mechanism. These basic facts of experience are the proper targets of introspection, if anything is. If introspection regularly fails to discern them correctly, it is not a reliable process.

xiii.

Descartes, I think, had it quite backwards when he said the mind—including especially current conscious experience—was better known than the outside world. The teetering stacks of paper around me, I'm quite sure of. My *visual experience* as I look at those papers, my *emotional experience* as I contemplate the mess, my *cognitive phenomenology* as I drift in thought, staring at them—of these, I'm much less certain. My experiences flee and scatter as I reflect. I feel unpracticed, poorly equipped with the tools, categories, and skills that might help me dissect them. They are gelatinous, disjointed, swift, shy, changeable. They are at once familiar and alien.

The tomato is stable. My visual experience as I look at the tomato shifts with each saccade, each blink, each observation of a blemish, each alteration of attention, with the adaptation of my eyes to lighting and color. My thoughts, my images, my itches, my pains—all bound away as I think about them, or remain only as self-conscious, interrupted versions of themselves. Nor can I hold them still, even as artificial specimens—as I reflect on one aspect of the experience, it alters and grows, or it crumbles. The unattended aspects undergo their own changes too. If outward things were so evasive, they'd also mystify and mislead.

I know better what's in the burrito I'm eating than I know my gustatory experience as I eat it. I know it has cheese. In describing my experience, I resort to saying, vaguely, that the burrito tastes 'cheesy,' without any very clear idea what this involves. Maybe, in fact, I'm just—or partly—inferring: The thing has cheese, so I must be having a taste experience of 'cheesiness.' Maybe also, if I know that the object I'm seeing is evenly red, I'll infer a visual experience of uniform 'redness' as I look at it. Or if I know that weeding is unpleasant work, I'll infer a negative emotion as I do it. Indeed, it can make great sense as a general strategy to start with judgments about plain, easily knowable facts of the outside world, then infer to what is more foreign and elusive, our *consciousness* as we experience that world.[31] I doubt we can fully disentangle such inferences from more 'genuinely introspective' processes.

Descartes thought, or is often portrayed as thinking, that we know our own experience first and most directly and then infer from that to the external world.[32] If that's right—if our judgments about the outside world, to be trustworthy, must be grounded in sound judgments about our experiences—then our epistemic situation is dire indeed. However, I see no reason to accept any such introspective foundationalism.[33] Indeed, I suspect the opposite is nearer the truth: Our judgments about the world to a large extent drive our judgments about our experience. Properly so, since the former are the more secure.

REFERENCES

Aristotle. *De Anima*, W. D. Ross, ed., (Oxford: Oxford University Press, 1961).

Armstrong, D. M., "Is Introspective Knowledge Incorrigible?" *Philosophical Review* 72 (1963): pp. 417–32.

Bar-On, Dorit. *Speaking My Mind* (Oxford: Oxford University Press, 2004).

Bayle, Pierre. *The Dictionary Historical and Critical of Mr. Peter Bayle*, Pierre des Maizeaux, trans., (London: Knapton et al., 1734–38 [1702]).

Bem, Daryl J. "Self-Perception Theory," *Advances in Experimental and Social Psychology* 6 (1972): pp. 1–62.

Berkeley, George, *A Treatise concerning the Principles of Human Knowledge* (1710), in *Principles, Dialogues, and Philosophical Correspondence*, C. M. Turbayne, ed., (New York: Macmillan, 1965), pp. 1–101.

Blackmore, Susan, "There Is No Stream of Consciousness," *Journal of Consciousness Studies* 9, nos. 5–6 (2002): pp. 17–28.

Boring, E. G. 1921. "The Stimulus Error." *American Journal of Psychology* 32: 449–71.

Brandstätter, Hermann, "Time Sampling Diary: An Ecological Approach to the Study of Emotion in Everyday Life Situations," in *Persons, Situations, and Emotions*, H. Brandstätter and A. Eliasz, ed., (Oxford: Oxford University Press, 2001), pp. 20–52.

Burge, Tyler, "Individualism and Self-Knowledge," *Journal of Philosophy* 85 (1988): pp. 649–63.

_____ "Our Entitlement to Self-Knowledge," *Proceedings of the Aristotelian Society* 96 (1996): pp. 91–116.

Chalmers, David J. *The Conscious Mind* (New York: Oxford University Press, 1996).

_____ "The Content and Epistemology of Phenomenal Belief," in *Consciousness: New Philosophical Essays*, Q. Smith and A. Jokic, ed., (Oxford: Clarendon, 2003), pp. 220–72.

Chisholm, Roderick M. *Perceiving* (Ithaca, NY: Cornell University Press, 1957).

Chuang Tzu. *Basic Writings*, B. Watson, trans., (New York: Columbia University Press, 1964).

Churchland, Paul M. "Reduction, Qualia, and the Direct Introspection of Brain States." *Journal of Philosophy* 82 (1985): pp. 8–28.

_____ *Matter and Consciousness* (Cambridge, MA: MIT Press, 1988), rev. ed.

Dennett, Daniel C. *Content and Consciousness* (New York: Humanities Press, 1969).

_____ *Consciousness Explained* (Boston: Little, Brown, and Co., 1991).

_____ "Surprise, Surprise," *Behavioral and Brain Sciences* 24 (2001): p. 982.

_____ "How Could I Be Wrong? How Wrong Could I Be?" *Journal of Consciousness Studies* 9, nos. 5–6 (2002): pp. 13–16.

Descartes, René, *Meditations on First Philosophy* (1641), in *The Philosophical Writings of Descartes*, J. Cottingham, R. Stoothoff, and D. Murdoch,

trans., (Cambridge: Cambridge University Press, 1984), pp. 1–62.

Dretske, Fred. *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995).

_____ *Perception, Knowledge, and Belief* (Cambridge: Cambridge University Press, 2000).

_____ "How Do You Know You are Not a Zombie?" in *Privileged Access*, B. Gertler ed., (Aldershot: Ashgate, 2003), pp. 1–13.

Ericsson, K. Anders, and Simon, Herbert A. *Protocol Analysis: Verbal Reports As Data* (Cambridge, MA: MIT Press, 1993 [1984]), rev. ed.

Gertler, Brie, "Introspecting Phenomenal States," *Philosophy and Phenomenological Research* 63 (2001): pp. 305–28.

_____, ed. *Privileged Access* (Aldershot: Ashgate, 2003).

Goldman, Alvin. *Epistemology and Cognition* (Cambridge, MA: Harvard University Press, 1986).

_____ "Epistemology and the Evidential Status of Introspective Reports," *Journal of Consciousness Studies* 11, nos. 7–8 (2004): pp. 1–16.

_____ *Simulating Minds* (New York: Oxford University Press, 2006).

Gordon, Robert M., "Simulation without Introspection or Inference from Me to You," in *Mental Simulation*, M. Davies and T. Stone, ed., (Oxford: Blackwell, 1995), pp. 53–67.

Haybron, Dan, "Do We Know How Happy We Are?" *Noûs* 41 (2007): pp. 394–428.

Hintikka, Jaakko, "Cogito Ergo Sum: Inference or Performance?" *Philosophical Review* 71 (1962): pp. 3–32.

Horgan, Terence, and Tienson, John, "The Intentionality of Phenomenology and the Phenomenology of Intentionality," in *Philosophy of Mind*, D. J. Chalmers, ed. (New York: Oxford University Press, 2002), pp. 520–33.

Horgan, Terence, John Tienson, and George Graham, "Internal-World Skepticism and the Self-Presentational Nature of Phenomenal Consciousness," in *Experience and Analysis: Proceedings of the 27th International Wittgenstein Symposium*, M. Reicher and J. Marek, eds., (Wien: ÖBV and HPT, 2005), pp. 191–207.

Huet, Pierre-Daniel, *Against Cartesian Philosophy* (1694), T. M. Lennon, ed. and trans., (New York: Humanity Books, 2003).

Hume, David, *An Enquiry concerning Human Understanding* (1748), in *Enquiries concerning Human Understanding and concerning the Principles of Morals*, L. A. Selby-Bigge and P. H. Nidditch, ed., (Oxford: Clarendon, 1975), pp. 5–165.

_____ *A Treatise of Human Nature* (1740), L. A. Selby-Bigge and P. H. Nidditch, eds., (Oxford: Clarendon, 1978).

Humphrey, George. *Thinking* (London: Methuen, 1951).

Hurlburt, Russell T. *Sampling Normal and Schizophrenic Inner Experience* (New York: Plenum, 1990).

Hurlburt, Russell T., and Schwitzgebel, Eric. *Describing Inner Experience? Proponent Meets Skeptic* (Cambridge, MA: MIT Press, 2007).

Jack, Anthony I., and Shallice, Tim, "Introspective Physicalism as an Approach to the Science of Consciousness," *Cognition* 79 (2001): pp. 161–96.

Jackson, Frank. *Perception* (Cambridge: Cambridge University Press, 1977).

James, William. *The Principles of Psychology* (1890) (Cambridge, MA: Harvard University Press, 1981).

Kornblith, Hilary, "What Is It Like to Be Me?" *Australasian Journal of Philosophy* 76 (1998): pp. 48–60.

Kriegel, Uriah, "The Same-Order Monitoring Theory of Consciousness," in *Self-Representational Approaches to Consciousness*, U. Kriegel and K. Williford, eds., (Cambridge, MA: MIT Press, 2006), pp. 143–70.

Lambie, John A., and Marcel, Anthony J., "Consciousness and the Varieties of Emotion Experience: A Theoretical Framework," *Psychological Review* 109 (2002): pp. 219–59.

Lawlor, Krista, "Knowing What One Believes," paper presented at University of California, Riverside, October 4, 2006.

Locke, John, *An Essay concerning Human Understanding* (1690), P. H. Nidditch. ed., (Oxford: Clarendon, 1975).

Lutz, Antoine, John D. Dunne, and Richard J. Davidson, "Meditation and the Neuroscience of Consciousness," in *Cambridge Handbook of Consciousness*, P. D. Zelazo, M. Moscovitch, and E. Thompson, eds., (Cambridge: Cambridge University Press, 2007), pp. 499–554.

Lycan, William G. *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996).

Mack, Arien, and Rock, Irvin. *Inattentional Blindness* (Cambridge, MA: MIT Press, 1998).

McGeer, Victoria, "Is 'Self-Knowledge' an Empirical Problem? Renegotiating the Space of Philosophical Explanation," *Journal of Philosophy* 93 (1996): pp. 483–515.

Montaigne, Michel de, *The Complete Essays of Montaigne* (1580), D. M. Frame, trans., (Stanford, CA: Stanford University Press, 1948).

Moran, Richard. *Authority and Estrangement* (Princeton, NJ: Princeton University Press, 2001).

Nichols, Shaun, and Stich, Stephen P. *Mindreading* (Oxford: Clarendon, 2003).

Nisbett, Richard E., and Wilson, Timothy DeCamp, "Telling More Than We Can Know: Verbal Reports on Mental Processes" *Psychological Review* 84 (1977): pp. 231–59.

Noë, Alva. *Action in Perception* (Cambridge, MA: MIT Press, 2004).

O'Regan, J. Kevin, "Solving the 'Real' Mysteries of Visual Perception: The World as an Outside Memory," *Canadian Journal of Psychology* 46 (1992): pp. 461–88.

Pitt, David, "The Phenomenology of Cognition; or, What Is It Like to Think That P?" *Philosophy and Phenomenological Research* 69 (2004): pp. 1–36.

Prinz, Jesse J. *Gut Reactions* (Oxford: Oxford University Press, 2004).

Rensink, Ronald A., J. Kevin O'Regan, and James J. Clark, "On the Failure to Detect Changes in Scenes

Across Brief Interruptions," *Visual Cognition* 7 (2000): pp. 127–45.

Robinson, William S., "Thoughts without Distinctive Non-Imagistic Phenomenology," *Philosophy and Phenomenological Research* 70 (2005): pp. 534–60.

Rosenthal, David M., "Two Concepts of Consciousness," *Philosophical Studies* 49 (1986): pp. 329–59.

Ryle, Gilbert. *The Concept of Mind* (New York: Barnes and Noble, 1949).

Sanches, Francisco, *That Nothing Is Known* (1581), E. Limbrick and D. F. S. Thomson, eds. and trans., (Cambridge: Cambridge University Press, 1988).

Schooler, Jonathan, and Schreiber, Charles A. "Experience, Meta-Consciousness, and the Paradox of Introspection," *Journal of Consciousness Studies* 11, nos. 7–8 (2004): pp. 17–39.

Schwitzgebel, Eric, "How Well Do We Know Our Own Conscious Experience? The Case of Visual Imagery," *Journal of Consciousness Studies* 9, nos. 5–6 (2002a): pp. 35–53.

_____ "Why Did We Think We Dreamed in Black and White?" *Studies in History and Philosophy of Science* 33 (2002b): pp. 649–60.

_____ "Introspective Training Apprehensively Defended: Reflections on Titchener's Lab Manual," *Journal of Consciousness Studies* 11, nos. 7–8 (2004): pp. 58–76.

_____ "Do Things Look Flat?" *Philosophy and Phenomenological Research* 72 (2006): pp. 589–99.

_____. 2007a. "Do You Have Constant Tactile Experience of Your Feet in Your Shoes? Or Is Experience Limited to What's in Attention?" *Journal of Consciousness Studies* 14 (no. 3): 5–35.

_____ "No Unchallengeable Epistemic Authority, of Any Sort, regarding Our Own Conscious Experience—Contra Dennett?" *Phenomenology and the Cognitive Sciences* 6 (2007b): pp. 107–13.

Schwitzgebel, Eric, and Gordon, Michael S. "How Well Do We Know Our Own Conscious Experience? The Case of Human Echolocation." *Philosophical Topics* 28 (2000): pp. 235–46.

Schwitzgebel, Eric, Changbing Huang, and Yifeng Zhou, "Do We Dream in Color? Cultural Variations and Skepticism," *Dreaming* 16 (2006): pp. 36–42.

Sextus Empiricus, *Outlines of Skepticism* (c. 200), J. Annas and J. Barnes, trans., (Cambridge: Cambridge University Press, 1994).

Shoemaker, Sydney, "Self-Knowledge and 'Inner Sense,'" *Philosophy and Phenomenological Research* 54 (1994): pp. 249–314.

Siewert, Charles P. *The Significance of Consciousness* (Princeton, NJ: Princeton University Press, 1998).

Skinner, B. F., "The Operational Analysis of Psychological Terms," *Psychological Review* 52 (1945): pp. 270–77.

Spener, Maja, "Phenomenal Adequacy and Introspective Evidence," unpublished manuscript, 2007.

Thomas, Nigel, "Are Theories of Imagery Theories of Imagination?" *Cognitive Science* 23 (1999): pp. 207–45.

Titchener, Edward Bradford. *A Primer of Psychology* (New York: Macmillan, 1899).

_____ *Experimental Psychology* (New York: Macmillan, 1901–5). (Note: The widely available 1971 reprint omits the instructor's part of the first volume.)

_____ *Lectures on the Experimental Psychology of the Thought-Processes* (New York: Macmillan, 1909).

_____ *A Text-Book of Psychology* (New York: Macmillan, 1910).

_____ "The Schema of Introspection," *American Journal of Psychology* 23 (1912): pp. 485–508.

Tye, Michael, "Representationalism and the Transparency of Experience," in *Privileged Access*, B. Gertler, ed., (Aldershot: Ashgate, 2003), pp. 31–43.

Unger, Peter. *Ignorance* (London: Clarendon, 1975).

Watson, John B., "Psychology as the Behaviorist Views It," *Psychological Review* 20 (1913): pp. 158–77.

Wilson, Robert A., "Intentionality and Phenomenology," *Pacific Philosophical Quarterly* 84 (2003): pp. 413–31.

NOTES

For helpful comments, criticism, and discussion, thanks to Donald Ainslie, Alvin Goldman, David Hunter, Tony Jack, Tori McGeer, Jennifer Nagel, Shaun Nichols, Gualtiero Piccinini, Josh Rust, Charles Siewert, Maja Spener (whose 2007 is similar in spirit to this essay), Aaron Zimmerman, and audiences at Washington University in St. Louis, Cal State Long Beach, University of Redlands, UC Santa Barbara, University of Toronto, and the Philosophy of Science Association.

1. For Descartes, see especially his "Second Meditation," (1984 [1641], p. 19). For Hume, see the first book of his *Treatise* (1978 [1739]), especially 1.4.2, 190, 212, and 1.4.5, 232. (Hume may change his mind in the *Enquiries*: see the first *Enquiry* [1975 {1748}], section 1, p. 13 and section 7, p. 60.) For Sextus, see *Outlines of Skepticism* (1994 [c. 200]), especially chapters 7 and 10. Pierre Bayle takes a similar position in the entry on Pyrrho in his *Dictionary* (1734–38 [1702], vol. 4, especially remark B, 654).
2. For Zhuangzi (third century BCE), see the second of his "Inner Chapters" (Chuang Tzu 1964). For Montaigne, see "Apology for Raymond Sebond" (1948 [1580]). Sanches' brief treatment of the understanding of the mind in *That Nothing Is Known* (1988 [1581], especially pp. 243–45 [57–59]) is at most only a partial exception to this tendency. So also is Unger 1975, 3, section 9), who seems to envision only the possibility of linguistic error about current experience and whose skepticism in this instance seems to turn principally upon an extremely demanding criterion for knowledge. Huet's *Against Cartesian Philosophy* (2003 [1694]) is nicely explicit in extending its skepticism to internal matters

of ongoing thought, though the examples and arguments differ considerably from mine here.

3. See especially his *Primer of Psychology* (Titchener 1899) and his *Experimental Psychology* (Titchener 1901–5). I discuss Titchener's views about introspective training at length in Schwitzgebel 2004. The Continental phenomenologists I find difficult to interpret on this point.
4. Consider Watson 1913; Skinner 1945; Ryle 1949; Bem 1972.
5. For example, Armstrong 1963; Churchland 1988—even Kornblith 1998, reading with a careful eye to distinguish error about current conscious experience from other sorts of error. See also, recently: Shoemaker 1994; Lycan 1996; Dretske 2000; Jack and Shallice 2001; Nichols and Stich 2003; Goldman 2004, 2006; Horgan, Tienson, and Graham 2005; and most of the essays collected in Gertler 2003, among many others. Gertler 2001 and Chalmers 2003 have recently attempted to revive restricted versions of (something like) infallibilism. Chalmers's infallibilism is so restricted I'm not sure how much useful substance remains. See section 9 for a discussion of the range and nature of infallible judgments.
6. For more on Dennett's granting people unchallengeable authority regarding their own experience, see Schwitzgebel 2007b.
7. I see no necessary conflict between the current view of introspection and views on which conscious experience involves a 'same order' (for example, Kriegel 2006) or 'higher-order' (for example, Rosenthal 1986; Lycan 1996) representation of the conscious state. Such views can allow—and to be plausible, I think they must allow—erroneous judgments of the sort to be discussed in this essay. For example, a non-conscious 'higher-order thought' that I am having experience *E* might conflict with a conscious judgment that I am not having experience *E*. Of course, only the conscious judgment is a reportable result of an introspective process.

   I do reject a strongly 'self-presentational' view of consciousness (as, perhaps, in Horgan, Tienson, and Graham 2005). The examples in the present essay, I think, reveal the implausibility of such an approach.

   Views characterizing us as constantly and effortlessly introspecting must either generate unreportable, nonconscious judgments, or they must in some other way differ, in mechanism or result, from the sort of self-conscious introspective efforts that are the topic of this essay and to which the term 'introspection' is here meant to refer.
8. Prinz 2004 helpfully reviews a variety of positions and evidence pertinent to them.
9. See, for example, Brandstätter 2001. It wouldn't surprise me in the least if positive mood even in studies such as this is considerably overreported.
10. Haybron 2007 presents an impressive array of evidence suggesting that we don't know how (un-)happy we are.
11. James 1981 (1890) and Lambie and Marcel 2002 may be a good place to start on this topic. In principle, of course, one could attempt to resolve such disputes by attributing vast individual differences in phenomenology to the participants, differences that perfectly mirror the divergences in their general claims; see sec. 11 for a discussion of this.
12. On skepticism about color in dreams see Schwitzgebel 2002b; Schwitzgebel, Huang, and Zhou 2006.
13. I take this argument to be in the spirit of Armstrong 1963. It needn't require that the phenomenology and the judgment be entirely 'distinct existences' in the sense Shoemaker 1994 criticizes, though of course it assumes that the one state is *possible* without the other. The only reason I see to reject such a possibility is a prior commitment to infallibilism.
14. For example, "Melanie" in Hurlburt and Schwitzgebel 2007.
15. See also Dennett 2001, 982.
16. In addition to this type of 'refrigerator light' error (Thomas 1999), an implicit analogy between visual experience and pictures or photographs may also sway us to over-ascribe detail in visual experience (see Noë 2004). Consider also Dennett 1969, pp. 139–41.
17. Among recent authors, Dennett 1991; O'Regan 1992; Mack and Rock 1998; Rensink, O'Regan, and Clark 2000; and Blackmore 2002 come to mind—though we differ somewhat in our positive views. Some of these authors believe we do not visually experience what we don't attend to. I mean to take no stand here on that particular question, which I explore in depth in Schwitzgebel 2007a.
18. The British empiricists (most famously, Locke 1975 (1690), Berkeley 1965 (1710), and Hume 1978 (1739) appear to have believed that conscious thought is always imagistic. So did many later introspective psychologists influenced by them (notably Titchener 1909, 1910), against advocates of 'imageless thought' (notably the 'Würzburg group,' whose work is reviewed in Humphrey 1951). Recent philosophers participating in the controversy include Siewert 1998, Horgan and Tienson 2002, Wilson 2003, Pitt 2004, and Robinson 2005. See also Aristotle *De Anima* (1961), 431a; Hurlburt and Schwitzgebel 2007, pp. 89–90.
19. These and related poll results were published at consc.net/neh/pollresults.html (accessed May 2005). I am inclined to read the disagreement between the 'no phenomenology of thought' and the 'imagery exhausts it' camps as a disagreement about terms or concepts rather than about phenomenology—a disagreement about whether having an image should count as 'thinking.' However, I see no similarly easy terminological explanation of the central dispute.

   As I recall (though this number is not recorded on the Web site), only two participants (Maja Spener and I) said they didn't know.
20. See Schwitzgebel and Gordon 2000; Schwitzgebel 2002a. See also Schwitzgebel 2006 for a discussion of divergent judgments about the experience of visual perspective, and Schwitzgebel 2007a for a discussion of our divergent judgments about whether we have a constant flow of peripheral experiences (of our feet in our shoes, the refrigerator hum, etc.).
21. I explore the possibility of classical introspective training, along the lines of early introspective psychology, in Schwitzgebel 2004 and the possibility of careful interviewing about randomly sampled experiences in Hurlburt and Schwitzgebel 2007. Schooler and Schreiber 2004 assesses the current scientific situation reasonably, if not quite as pessimistically. Very recently, there has been some promising work on meditation: see Lutz, Dunne, and Davidson 2007.
22. Compare Hintikka 1962; Burge 1988, 1996.

23. But see Chalmers 1996 and Dretske 2003 on the possibility that we could be experienceless 'zombies' without knowing it. Both Chalmers and Dretske think we do know that we are conscious, but that it's not straightforward to see *how* we know that.
24. Compare Chisholm 1957; Jackson 1977. Naturally, ordinary and philosophical usage of 'appears' is rather more complex than this simple portrayal suggests if one looks at the details; but I don't think that affects the basic observation of this section.
25. See Moran 2001 and Bar-On 2004 for versions of this story.
26. For more on mistakes in the introspection of nonobvious or fictional illusions, see Schwitzgebel 2004.
27. For example, James 1981 (1890), p. 189; Titchener 1912, p. 491; Hurlburt 1990, chapter 2.
28. Epistemologists often define 'reliability' so that only the first type of failure counts as a failure of reliability (for example, Goldman 1986, who calls the second sort of failure a lack of 'power'). It's a semantic issue, but I think ordinary language is on my side.
29. See Gordon 1995; McGeer 1996; Moran 2001; Bar-On 2004; Lawlor 2006.
30. Ericsson and Simon 1993 (1984) are optimistic about the accuracy of descriptions of one's thought processes when one 'thinks aloud,' expressing the thought concurrently with having it. They are considerably less optimistic about retrospective reports if the subject is not primed and trained in advance to express and reflect on her thoughts as they occur.

    Burge 1996 argues that, to be successful, 'critical reasoning' requires knowledge of recently past thought contents. But I doubt much of our reasoning is 'critical' in the relevant sense. (Usually, it is spontaneous and un-self-reflective; often it is entirely hidden.) Nor is it clear that, when we try to reflect critically on our stream of reasoning, we are reliably successful in doing so.
31. Titchener thinks this strategy common among untutored introspectors, and he repeatedly warns against it as 'stimulus error' or 'R-error': Titchener 1901–5; Boring 1921. This strategy bears some relation to the strategy that 'transparency theorists' such as Dretske 1995, 2000 and Tye 2003 think we always use in reaching judgments about our experience (though they hardly think of experience as 'elusive').
32. Whether this is the best interpretation of Descartes, I am uncertain. My impression is that Descartes is not entirely clear on this point, and sympathetic interpretations of him shift with the mood of the times. The view is also associated with Locke 1975 (1690).
33. Of course, if it were possible to draw a clear line between the trustworthy and untrustworthy introspective judgments, then maybe a version of introspective foundationalism could be salvaged. I'm not optimistic that such a line could be drawn, or that, if it were, enough trustworthy introspective judgments would remain to be of much use.

What You Can't Expect When You're Expecting

L. A. Paul

It seems natural to choose whether to have a child by reflecting on what it would be like to have one. I argue that choosing on this basis is not rational, raising general questions about our ordinary conception of how to make this life-changing decision.[1]

1. Deciding Whether to Start a Family

Scenario: You have no children. However, you have reached a point in your life when you are personally, financially and physically able to have a child.[2] You sit down and think about whether you want to have a child of your very own. You discuss it with your partner and contemplate your options, carefully reflecting on the choice by assessing what you think it would be like for you to have a child of your very own and comparing this to what you think it would be like to remain childless. After careful consideration, you choose one of these options:

For: You decide to have a child.
Against: You decide to remain childless.

The way you went about making your choice seems perfectly apt. It follows the cultural norms of our society, where couples are encouraged to think carefully and clearly about what they want before deciding that they want to start a family. Many prospective parents decide to have a baby because they have a deep desire

From *Res Philosophica* 92:1–23, 2015.

to have children based on the (perhaps inarticulate) sense that having a child will help them to live a fuller, happier, and somehow more complete life.[3] While many people recognize that an individual's choice to have a child has important external implications, the decision is thought to necessarily involve an intimate, personal component, and so it is a decision that is best made from the personal standpoints of prospective parents.[4] Guides for prospective parents often suggest that people ask themselves if having a baby will enhance an already happy life, and encourage prospective parents to reflect on, for example, how they see themselves in five and ten years' time, whether they feel ready to care for and nurture the human being they've created, whether they think they'd be a happy and content mother (or father), whether having a baby of their own would make life more meaningful, whether they are ready for the tradeoffs that come with being a parent, whether they desire to continue with their current career plans or other personal projects, and so on.[5]

This assessment of one's prospects and plans for the future is a culturally important part of the procedure that one is supposed to undergo before attempting to become pregnant. Since (in the usual case) the parents assume primary responsibility for the child they create, it seems appropriate to frame the decision in terms of making a personal choice, one that carefully weighs the value of one's future experiences.[6] People often frame the decision this way when they make this choice, and more importantly for my purpose here, we are (culturally speaking) supposed to frame the decision this way. Given the magnitude of the responsibilities we are considering taking on, we are *supposed* to think carefully about the personal implications of the choice. Many choose to have a child. Many prefer to remain childless.

2. Decision Theory: A Normative Model

When we make a choice to do something, we make a decision: we consider various things we might do and then choose to do one of them, and decision theory provides the best account of rational decision-making. Ideal agents in ideal circumstances make choices rationally by conforming to the models of an idealized decision theory. To make a choice rationally, we first determine the possible outcomes of each act we might perform. After we have the space of possible outcomes, we determine the value (or utility) of each outcome, and determine the probability of each outcome's occurring given the performance of the act. We then calculate the expected value of each outcome by multiplying the value of the outcome by its probability, and choose to perform the act with the outcome or outcomes with the highest overall expected value.

Now, decisions made by real agents in real-world circumstances do not conform to this standard model. Ordinary reasoners may be imperfect reasoners; their reasoning may only imperfectly conform to the way an ideal rational being would reason, and their assessments of the values of the outcomes may only imperfectly conform to their actual values. A more realistic version of a decision-theoretic approach, that is, what I'll call a *normative* decision theory, can capture norms for ordinary successful reasoning. If we can glean approximate values for our outcomes and apply the right decision theoretic rules, we can conform to the ordinary standard for rational decision-making. Decisions made by ordinary people can be rational if they conform to the realistic standards set by a normative decision theory, where such standards make allowances for a certain amount of approximation, ignorance, uncertainty, and mistaken beliefs.[7]

For example, when considering an outcome, perhaps we can do no better than glean its approximate expected value. After all, it is probably impossible for a person to calculate the expected value of each outcome with precision. And perhaps we do not know about all the possible outcomes. But we can approximate a rational choice by choosing between approximate expected values of the relevant or the most important outcomes. A normative decision theory describes the range and combination of rules and standards that agents must meet for their decisions to be rational, normatively speaking. It thus provides a *normative model* that real agents can conform to so that their decisions are rational by our lights.[8] In this paper, I will assume that we want to meet the standard for normative rationality when we make the decision of whether or not to have a child.

In any non-ideal case, complicating features may be present. For example, sometimes outcomes have equal expected values. Then no unique act is the rational one to choose. Sometimes expected values are metaphysically indeterminate. Then it is metaphysically indeterminate which act is the rational one to choose. Or perhaps we cannot adequately partition the

space of possible outcomes. Etc. For simplicity, I assume that such features are not present in *Scenario*. In particular, I assume that we can partition the space of relevant possibilities into a set of suitably fine-grained, exclusive and exhaustive propositions describing each relevant outcome.

In *Scenario*, the acts in question are either having one's own child or not having one's own child. The decision is the choice between whether to have a child or whether to remain childless. The outcomes of either act are its effects, which have dramatic emotional, mental and physical consequences. The dramatic effects follow the act of not having a child as much the act of having one: for example, not having a child means that you'll have very different experiences from ones you'd have had if you had a child, and has follow-on effects, such as the fact that you'd have significantly fewer financial costs for at least eighteen years following the date from when the omission can be said to 'obtain.'

The primary concern in *Scenario* is with the value of the outcome 'for the agent,' where this describes the value of the outcome brought about by the agent, centering on the outcome that involves the agent's perspective or point of view, that is, on the subjective value of what it is like to be the person who made the choice. In particular, the agent in *Scenario* is concerned with phenomenal outcomes that involve what it's like for her to have her own child. Since what it is like to be the agent includes what it is like to have her beliefs, desires, emotions, dispositions, and to perform subsequent acts, in *Scenario* the relevant outcomes include what it is like to have these additional effects and their attendant consequences as part of what it is like for her to have her child.

When choosing between *For* or *Against*, you compare the overall expected values of the outcomes of each act. Since we are concerned here with ordinary decision-making, we use a normative model to guide our choice, allowing for approximation and estimation in place of perfect precision. To choose rationally, given our normative model, you determine the approximate value of each relevant outcome, you determine the approximate probability of each of these outcomes actually obtaining, and then use this information to estimate the expected value of each act. After estimating the expected value of each act, you choose the act that brings about the outcome with the highest estimated expected value.

In the case where you have a child, the relevant outcomes are phenomenal outcomes concerning what it is like for you to have your child, including what it is like to have the beliefs, desires, emotions and dispositions that result, directly and indirectly, from having your own child. Thus, the relevant values are determined by what it is like for you to have your child, including what it is like to have the beliefs, desires, emotions and dispositions that result, directly and indirectly, from having your own child. (I will sometimes call these values 'phenomenal values': they are values of being in mental states with a phenomenal 'what it's like' character.) In the case where you remain childless, the relevant outcomes are phenomenal outcomes involving what it is like for you to experience the effects of remaining childless, and thus the relevant values depend on what it is like for you to experience childlessness. In other words, the value of your act in *Scenario*, given the way the choice is made, depends largely on the phenomenal character of the mental states that result from it. This is neither surprising nor unusual from a commonsensical point of view.

Of course, having a child or not having a child will have value with respect to plenty of other things, such as the local demographic and the environment. However, the primary focus here is on an agent who is trying to decide, largely independently of these external or impersonal factors, whether she wants to have a child of her own. In this case, the value of what it is like for the agent plays the central role, if not the only role, in the decision to procreate. That said, the value of the choice is also affected if we assess the wider scope of the value of the act, since even in cases with a wider purview, the value of what it is like for the agent to have her own child must be evaluated in order to determine the overall expected value of her choice. For instance, you might choose to have a child because you desire to have some of your DNA transmitted to future generations. But the value of satisfying this desire must be weighed against other outcomes. If, say, the value of what it was like for you to have your own child was sufficiently positive or sufficiently negative, it could swamp the value of satisfying your desire to leave a genetic imprint.

3. What Experience Teaches

All of this might seem perfectly straightforward and unexceptionable. But there is a problem lurking beneath the surface. To see it, begin by reflecting on an interesting fact about 'what it's

like' knowledge, such as knowledge of what it's like to see red. The interesting fact is that this sort of knowledge, that is, knowing what it's like, can (practically speaking) *only* be had via experience.

Frank Jackson developed a famous thought experiment to make this point. His example features black-and-white Mary, a brilliant neuroscientist, who is locked in a colorless cell from birth. Mary has never experienced color. Now, she knows all the facts in a complete physics (and other sciences), including all the causal and relational facts and functional roles consequent on knowing these facts, and including all the scientific facts about light, the human eye's response to light with wavelengths between 600 and 800 nanometers and any relevant neuroscience. Yet, when she has her first experience of red, she learns something new: she learns what it is like to see red.

> Mary is confined to a black-and-white room, is educated through black-and-white books and through lectures relayed on black-and-white television. In this way she learns everything there is to know about the physical nature of the world. . . . It seems, however, that Mary does not know all there is to know. For when she is let out of the black-and-white room or given a color television, she will learn what it is like to see something red. . . .' (Jackson 1986, p. 291)

As Jackson points out, when Mary leaves her cell for the first time, she has a radically new experience: she experiences redness for the first time, and from this experience, and this experience alone, she knows what it is like to see red.

Because of Mary's lack of experience, before she leaves her black-and-white cell, she lacks a certain kind of knowledge. Perhaps that knowledge is knowledge of a physical fact. Perhaps that knowledge involves a lack of a certain kind of ability or know-how. Perhaps it's knowing an old fact in a new way. Or perhaps, after leaving her room, she knows a new fact of some other sort.[9] None of that matters here.[10] The lesson for us is simply that, before she leaves her cell, black-and-white Mary is in an impoverished epistemic position. Until she actually has the experience of seeing red, she cannot know what it is like to see red.

An important feature of this example relies on the fact that, given Mary's exclusively black and white experiences, the experience of seeing red is unique and distinctive for her. Before she leaves her room, she cannot project forward to get a sense of what it will be like for her to see red, since she cannot project from what she knows about her other experiences to know what it is like to see color. As the example is described, then, before she leaves the room, her previous experience is not projectable in a way that will give her information about what it is like to see red. As a result, when she leaves her room and sees red for the first time, her experience is *epistemically transformative*.

Now let's restrict Mary's epistemic situation a little more than it was in Jackson's thought experiment. Before she leaves her room, because she doesn't know what it is like to see red, or indeed what it is like to see any sort of color at all, she also doesn't know what feelings and thoughts she'll experience as the result of seeing red.[11] And so she doesn't know whether it'll be her favorite color, or whether it'll be fun to see red, or whether it'll be joyous to see red, or frightening to see it, or whatever. And even if she could know, say, that she would find seeing red frightening, she wouldn't know how phenomenologically intense this experience would be.

For our purposes, Mary's impoverished epistemic situation means, first, that since Mary doesn't know how it'll phenomenally feel to see red before she sees it, she also doesn't know what emotions, beliefs, desires, and dispositions will be caused by what it's like for her to see red. Maybe she'll feel joy and elation. Or maybe she'll feel fear and despair. And so on. Second, because she doesn't know what emotions, beliefs, desires, and dispositions will be caused by her experience of seeing red, she doesn't know what it'll be like to have the set of emotions, beliefs, desires, and dispositions that are caused by her experience of seeing red, simply because she has no guide to which set she'll actually have. And third: she doesn't know what it'll be like to have any of the phenomenal-redness-involving emotions, beliefs, desires, and dispositions that will be caused by her experience of seeing red. Even if she could somehow know that she'll feel joy upon seeing red, she doesn't know what it will be like to feel-joy-while-seeing-redness until she has the experience of seeing red. And these are all ways of saying that, before she leaves her cell, she cannot know the value of what it'll be like for her to see red.

This means that, when Mary chooses to leave her black-and-white cell, thus choosing to undergo an epistemically transformative experience, she faces a deep subjective unpredictability about the future. She doesn't know, and she cannot know, the values of the relevant phenomenal outcomes of her choice.

4. The Transformative Experience of Having a Child

A person who is choosing whether to become a parent, before she has a child, is in an epistemic situation just like that of black-and-white Mary before she leaves her cell. Just like Mary, she is epistemically impoverished, because she does not know what it is like to have a child of her very own.

Why is she epistemically impoverished? At least in the normal case, one has a uniquely new experience when one has one's first child. Before someone becomes a parent, she has never experienced the unique state of seeing and touching her newborn child. She has never experienced the full compendium of the extremely intense series of beliefs, emotions, physical exhaustion and emotional intensity that attends the carrying, birth, presentation, and care of her very own child, and hence she does not know what it is like to have these experiences.

Moreover, since having one's own child is unlike any other human experience, before she has had the experience of seeing and touching her newborn child, not only does she not know what it is like to have her child, she cannot know.[12] Like the experience of seeing color for the first time, the experience of having a child is not projectable. All of this means that having a child is epistemically transformative.

Now, having a child is not just a radically new epistemic experience, it is, for many people, a life-changing experience. That is, the experience may be both epistemically transformative and *personally transformative*: it may change your personal phenomenology in deep and far-reaching ways. A personally transformative experience radically changes what it is like to be you, perhaps by replacing your core preferences with very different ones.[13] For most people, having a child is transformative in both ways: it is an epistemically transformative experience that is also personally transformative.

Why do parents experience such dramatic phenomenological changes? It is a normal reaction to the intense series of new experiences that one has when one has a child of one's own. This is most obvious when the parent in question is the mother. The intensity and uniqueness of the extended act of carrying the child, the physicality of giving birth, the recognition of the new fact of the existence of one's very own child, and the exertion involved in caring for a newborn results in a dramatic change in one's physical, emotional and mental states. The experiences are also very intense for involved fathers. It is common for fathers to date their changed phenomenal state from the moment they saw or held their newborn.

Perhaps the primary basis for the radical change in phenomenology in both parents is the simple fact that the content of the state of *seeing and touching your own newborn child* can carry with it an epistemically unique and personally transformative phenomenological character.[14] This may be the source of why this experience is both epistemically and personally transformative.

There are probably attendant biological reasons for the phenomenological change in parents: when producing, breastfeeding and caring for a child, mothers experience enormous hormonal and other biological changes, and new fathers also undergo significant hormonal changes. Fans of evolutionary biology will hold that there is a biological mandate for the physiological changes in the parents that underlie the felt attachment to one's offspring. In any case, whether the primary basis for one's new phenomenology is simply the experience of producing, seeing, and touching your newborn child, or whether it is being in some new biological state, or whether it is a more extended and complex series of experiences, the parent has an experience he or she has never had before—an experience with an epistemically unique phenomenal character, and moreover, one which can also be personally transformative.[15]

The combination of the epistemically and personally transformative experience of having one's own child brings with it profound changes in other epistemic states. In particular, because you cannot know what it is like to have your own child before you've had her, you also cannot know what emotions, beliefs, desires, and dispositions will be caused by what it's like to have her. Maybe you'll feel joy and elation when she is born. Or maybe you'll feel anger and despair (many parents experience postnatal depression). And so on. Moreover, you can't know what it'll be like to have the particular emotions, beliefs, desires, and dispositions that are caused by your experience of having your child. As a result, if you have a child, and if your experience is both epistemically and personally transformative, many of your epistemic states will change in subjectively unprojectable ways, and many of these changes will be profound changes.

5. Choosing the Ordinary Way Is Not Rational

Recall the normative model for ordinary decision-making given in section 2. You, as a normatively rational agent, are supposed to deliberate between acts: you determine the relevant outcomes of each act, the approximate probability of these outcomes, the approximate value of these outcomes, and then estimate the overall expected value of each act. After estimating the expected value of each act, you choose the act that has the highest expected value.

The lurking problem I alluded to in section 3 comes from the fact that the normative model requires one to determine values of outcomes. And, in fact, *any* standard decision-theoretic model requires one to determine values, at least approximate ones, of outcomes. The problem surfaces when we realize that, first, we want to make the decision based on the phenomenal outcome, that is, based on what we think it will be like to have a child. And second, that if our choice involves an outcome that is epistemically transformative, we cannot know the value of this outcome before we experience it. And if we cannot determine the value of the relevant outcome, we are in the same epistemic position as the agent who, because he doesn't know what the prize will be, cannot rationally determine the utility of winning the lottery (Weirich 2004, p. 65).

Recall Mary in her black-and-white cell. Imagine that she is trying to decide whether she wants to leave her cell for the first time. As we saw, Mary doesn't know what it will be like to see color. In addition to its being a certain way to see red, maybe it will be terrifying and overwhelming to see color after living in black and white for so long. Maybe the particular fear created by seeing redness will be mind-numbingly awful and paralyzing. Or maybe seeing red for the first time will be blissfully wonderful. She just doesn't know. As I noted above, this means Mary doesn't know what values to assign to the phenomenal states that are the outcomes of her choice to leave her cell. If she cannot rationally determine the values of the relevant outcomes, she cannot use normative decision theory to make a rational choice. (And if she assigns values to these phenomenal states anyway, she is making an unacceptable mistake, for if she cannot know their values, there are no rationally acceptable values she can assign.) Either the decision theoretic model does not apply, because there is no value known for the relevant outcome, or the value she assigns to the outcome is based on an unacceptable belief about what the value should be, and a decision based on an unacceptable belief is not rational.[16]

The very same problem arises in *Scenario*. Here, you are deciding whether to have a child based on the expected value of the act for you and your partner. You think about what it would be like to have a child, how it will affect you and your partner, and how it will affect the other parts of your life, and you decide on the outcome with the best overall effects, where 'best overall effects' is short for 'effects that maximize expected value.' Even if the contemplation is not as detailed or precise as the perfect rational agent could make it, an approximation of this approach embodies our ordinary way of trying to take a clear-headed, normatively rational approach to this extremely important decision.

The trouble comes from the fact that, because having one's first child is epistemically transformative, one cannot determine the value of what it's like to have one's own child before actually having her. This means that the subjective unpredictability attending the act of having one's first child makes the story about family planning into little more than pleasant fiction. Because you cannot know the value of the relevant outcome, there is no rationally acceptable value you can assign to it. The problem is not that a prospective parent can only grasp the approximate values of the outcomes of her act, for then, at least, she might have some hope of meeting our norms for ordinary decision-making. The problem is that she cannot determine the values with any degree of accuracy at all.

As a result, no matter which option in *Scenario* you choose, your decision is not even an approximation of a normatively rational act. It is impossible for you to follow the decision procedure in *Scenario* and choose *For* in a way that is consistent with the ordinary standard for rational decision-making. It is also impossible for you to follow the decision procedure in *Scenario* and choose *Against* in a way that is consistent with the ordinary standard for rational decision-making. Arguably, ordinary rationality does not even *permit* making either choice. Generalizing this, you cannot use our ordinary, phenomenal-based, normative decision procedure to rationally make one of the biggest decisions of your life. You cannot use this procedure to rationally choose to have

a child, nor to rationally choose to remain childless.

Distinguishing between evidential and causal probability does not help: it is not rational to choose either option whether we understand your decision as one based on evidence or as one based on a judgment about the causal efficacy of the act. Finally, even a distinction between practical rationality and theoretical rationality will not help: your choice in *Scenario* is neither theoretically nor practically rational in the intended sense.[17]

It should be obvious that, in this discussion, I am abstracting from any moral considerations that might affect the choice to have or not to have children, and I am not taking a position on the nature of moral deliberation—i.e., whether it is a form of rational deliberation, and whether its aim is to maximize value. I am starting from what I take to be our predominant cultural paradigm of how to consider the question of whether to have or not to have a child. According to that paradigm, we are to approach this decision as a personal matter where what is at stake is our own expected happiness and a sort of personal self-realization.[18]

And so we find a conflict between the ordinary way we are supposed to make the decision to have a child and the fact that having one's own child is an epistemically transformative experience. This conflict is interesting precisely because the decision to have a child may also be personally transformative. When a decision involves an outcome that is epistemically transformative for the decision-maker, she cannot rationally assign a value to the outcome until she has experienced the outcome. When that outcome may also be personally transformative for the decision-maker, the conflict matters—for she needs to make a big decision, a possibly self-transformative decision, and she cannot conform to ordinary or 'folk' norms for rational decision-making when doing so.

6. Objections

My conclusion is controversial. The remainder of the paper will discuss some objections.

6.1 Subjective Ability

Perhaps you think that you can know what it's like to have a child, even though you've never had one, because you can read or listen to the testimony of what it was like for others. You are wrong.

> If you want to know what some new and different experience is like, you can learn it by going out and really *having* that experience. You can't learn it by being told about the experience, however thorough your lessons might be. . . . You may have tasted Vegemite, that famous Australian substance; and I never have. So you may know what it's like to taste Vegemite. I don't, and unless I taste Vegemite (what, and spoil a good example!) I never will. (Lewis 1990, p. 292)

The experience of having a child is exactly the sort of epistemically unique, epistemically new experience that Lewis is referring to.[19] Having one's first child and tasting Vegemite for the first time are both epistemically transformative (though tasting Vegemite is rarely personally transformative, unless you are an Australian who has been away from home for a long time).

Being around other people's children isn't enough to learn about what it will be like in your own case. The resemblance simply isn't close enough in the relevant respects. Babysitting for other children, having nieces and nephews or much younger siblings—all of these can be wonderful (or horrible) experiences, but they are different in kind from having a child of your very own, perhaps roughly analogous to the way an original artwork has aesthetic value partly because of its origins. (Thus the various memes about 'other people's children,' including those about how one can dislike other people's children while loving one's own, about how adopting a child 'isn't the same' as having one,[20] etc.) Experience with other peoples' children might teach you about what it is like to hold a baby, to change diapers or hold a bottle, but not what it is like to create, carry, give birth to and raise a child *of your very own*. This is obvious even if we discount the conceptual or indexical basis for the uniqueness of the experience, for as I pointed out above, there are purely biological causes that may be sufficient for its uniqueness: the hormonal reactions and other biological responses that stem from physically growing, carrying and giving birth to your own child (mutatis mutandis for fathers). One simply does not get this biological response from babysitting one's niece or changing one's nephew's dirty diaper.

You might think that having a description of what it's like to have a child will tell you what you need to know if it tells you about other experiences that closely resemble the new experience. But it doesn't, at least if you haven't experienced anything that closely resembles

the experience, such as already having a child of your own. Lewis 1990, pp. 265–66, points out that even if one can be told that the taste of Vegemite somewhat resembles Marmite, unless one has tasted Marmite, this misses the point. Without the relevant experience, no amount of information about resemblances will help.

The claim that having a child is epistemically transformative does not entail that, if you ascribe a value to what it will be like for you to have a child before you've actually had a child, the value you ascribe will be incorrect. You might get lucky. You might ascribe a value that, once you have the child, turns out to be reasonably close to the actual value. But this doesn't mean that it was rationally acceptable for you to ascribe this value before you could know what it was going to be. It was not rationally acceptable, for you could not know the value before you'd had the experience.[21]

Back to Mary in her colorless cell: Mary might guess that the experience of seeing color for the first time will be stressful and frightening. When she leaves her cell, she might indeed find her experiences of redness to be stressful and frightening. Or Mary might guess that the experience of seeing color for the first time will be fulfilling and satisfying. When she leaves her cell, she might indeed find her experiences of redness to be fulfilling and satisfying. But none of this entails that she was able to know what it would be like for her to experience redness before she actually experienced it, and so none of this entails that it was rationally acceptable for Mary to assign these values before she left her cell.

Can there really be anyone who would grant that the relatively mundane experience of tasting Vegemite for the first time is epistemically transformative, while *denying* that growing, carrying, giving birth to, and raising one's first child is epistemically transformative? If you grant that epistemically transformative experiences are possible at all, you should grant that having your first child is one of them.

6.2 Alternative Decision Procedures

The normative model captures the structure of an ordinary decision-making process. Many people, myself included, take the normative model (or close variations thereof) to provide the most natural framework for decision-making in this particular context, even if it gives us unsatisfactory results. However, it is well-known that decision-making under ignorance creates special problems for agents, and models for decision-making under ignorance have been developed for agents to use.[22] How does this fact affect my argument?

In a nutshell: it doesn't. Our option is to replace the simple version of the normative model with a different version, one which would apply under epistemically impoverished circumstances. This might seem like the obvious way to approach the problem. After all, the real world is messy, and as I discussed in §2, the difficulty of fitting the pristine, clear and precise models of decision theory with the murky viewpoints of actual agents is well-known. Can we accommodate decisions involving epistemically transformative experiences by using special models for decision-making under ignorance?

No. The same problem that arose for our simple normative model arises with these special models, for it is a condition of application for all such models that we are able to legitimately determine the values (or utilities), at least approximate ones, of the relevant outcomes of the act.[23] In the most common models for decision under ignorance, the models specify the values of the outcomes of the act, but—representing agent ignorance—no probabilities are determined. Just as with our original normative model, your choice to have your own child is based on your phenomenal preferences, so to use these decision theoretic models, you have to be able to determine the approximate values of the phenomenal outcomes, outcomes including *what it is like for you to have your own child.* But because you do not know what it is like to have your own child, you lack the relevant phenomenal knowledge you need in order to rationally determine these values.

For example, a simple model for decision-making under ignorance could use the 'maximin' rule for making decisions. When 'maximining' the agent decides conservatively, that is, makes a safe bet, with the objective of minimizing bad results. To use this decision procedure, we first determine the desirability and undesirability of each relevant outcome. Then we choose the act whose worst outcome has the highest desirability relative to the worst outcomes of all the acts under consideration, that is we, choose the act with the 'least bad' outcome. A different, more optimistic model uses a version of the 'maximax' rule: calculate the value of each relevant outcome, and then simply choose the outcome that has the highest value. That is, we 'maximax' by choosing the

act whose best outcome is the most desirable outcome. Either approach allows for rational decision-making under ignorance.

To apply these models, we determine the values of outcomes and then apply a decision rule. The appropriate decision rule depends on the context, which includes the agent's circumstances and dispositions. If, for example, you are choosing from a range of unfamiliar dishes at a new restaurant somewhere in the Midwest, you might wish to employ the maximin rule, selecting the simply prepared steak instead of the interesting, but unusually flavored, seafood dish. Here, outcomes include having a decent steak, having a delicious seafood dish, or having a disturbingly chewy, unpleasantly fishy evening meal. On the other hand, if the restaurant has enough Michelin stars, you might decide to throw caution to the winds and employ maximax reasoning to go for the Aguachile de Pulpo y Calamar after all.

But what if you are visiting Australia for the first time, and need to choose between having toast with orange marmalade and toast with Vegemite? If you've never had Vegemite, nor anything resembling it (such as Marmite), and you want to choose based on what it will be like for you to taste Vegemite, you are out of luck.[24] Neither maximin nor maximax will work for you. In the Midwestern restaurant, you chose between outcomes that resembled what you'd experienced in the past (a decent steak, good seafood, bad seafood), and so you were able to assign values to them. But in a case where you really don't know what it's like to taste the menu item, you can't use maximin, or maximax, or any other decision-under-ignorance rule to rationally make a decision based on what you think it will taste like. You just don't have enough information to deploy the model.[25] You might be able to rationally make your menu choice on another basis, say, where you regard the choice merely as a fun, low-stakes gamble, but a decision on that basis is not analogous to the phenomenally-based decision to have a child.

You might think, hang on, we can just parse the range of outcomes so that they are described as outcomes like 'Vegemite tastes delicious,' and 'Vegemite tastes disgusting.'[26] But simply adding terms like 'delicious' or 'disgusting' to the description of the outcome won't give you the information about values that you need. Intuitively speaking, you need to know more in order to assign them values. You need to know how phenomenally intense the state described by 'Vegemite tastes delicious' and how phenomenally intense the state described by 'Vegemite tastes disgusting' is, and you need experience in order to know this.[27]

We find ourselves with the very same problem in *Scenario*. No standard model of decision under ignorance is available to the prospective parent who chooses based on what she thinks it will be like to be a parent, for, just as in the Vegemite case, she cannot determine the values of the relevant outcomes. As a result, the models don't apply.

Now, of course, I am assuming various constraints here: it isn't *metaphysically impossible* to determine the values of the outcomes. It is simply epistemically impossible given very reasonable and appropriate real-world constraints. For example, if you had a perfect physical duplicate who underwent the experience of having a child and then told you how to assign values to the outcomes for your version of the experience, you could employ a decision-theoretic model. This sort of pretend scenario, and various other sci-fi alternatives we might be able to dream up, are obviously irrelevant in this context.

There is another issue here that should be raised: not only is the phenomenal outcome *what it's like to have your own child* a relevant outcome of your choice, it's an outcome whose value might *swamp* the other outcomes. In other words, even if other outcomes are relevant, the value of the phenomenal outcome, when it occurs, might be so positive or so negative that none of the values of the other relevant outcomes matter.[28]

Now, we need not take the fact that normative decision theoretic models don't work well for the case of having children as a criticism of decision theory, for sophisticated decision theorists often think of decision theory as a useful evaluative tool, not as a method one should use to determine, in practical circumstances, what sort of deliberation is rational.[29] The point being made here is that you cannot rationally decide to have a child based on what you think it will be like for you to have a child, and debates about how to make this important life choice should reflect this fact.

6.3 Eliminate the Subjectivity in the Decision Procedure

The source of the problem is the epistemically transformative nature of the experience of having one's child. One way to circumvent this

problem is by dispensing with projectability, that is, ignoring your own personal preferences when you choose. You can change the decision procedure and choose to have a child based *solely* on the assumption that anyone who has a child is more likely to end up in a class of individuals who maximize their overall utility, ignoring your own personal beliefs, desires and other phenomenal projections about the future.

Let's consider this possibility. After choosing, you could end up in one of four different classes. The class of individuals for whom, after having a child, the overall value of having a child is higher than it would have been if they had remained childless, is *Lucky Parents*. The class of individuals for whom, after having a child, the overall value of having a child is lower than it would have been if they had remained childless, is *Unlucky Parents*. The class of individuals for whom, having decided to not have a child, the overall value of the choice to be childless is higher than it would have been if they had had a child, is *Lucky Child-frees*. Finally, the class I'll label *Unlucky Child-frees* is the class of individuals for whom, having decided to be childless, the overall value of the choice to not have a child is lower than it would have been if they had had a child.

Now if Lucky Parents is much larger than Unlucky Parents, and Unlucky Child-frees is much larger than Lucky Child-frees, it might seem rational to choose to have a child, simply because you think, given the numbers, if you have a child you are far more likely to be in Lucky Parents than in Unlucky Parents, and you successfully avoid being classed in Unlucky Child-frees. And indeed, many people seem to assume something like the claim that Lucky Parents is much larger than Unlucky Parents. They also seem to assume that Unlucky Child-frees is much larger than Lucky Child-frees: they assume that people increase their happiness and well-being by having children and that childless people decrease their well-being (and as a result are unhappy or unfulfilled) because they do not have children of their own.

However, current empirical evidence suggests that this assumption is false. While the highs seem to be higher for parents, the lows seem to be lower, and many measures suggest that parents with children in the home have, on average, a lower level of overall life satisfaction.[30] Moreover, individuals who have never had children report similar levels of life satisfaction as individuals with grown children who have left home (Simon 2008; Evenson and Simon 2005). A recent analysis of survey data covering a wide range of the empirical results concerning parenthood indicates that *no* group of parents, including those whose children have grown and left home, where those groups are determined by standard sociological classifications such as income, marital status, gender, race, education, and mental health, report higher levels of overall emotional well-being than non-parents (Simon 2008; Evenson and Simon 2005).[31] Psychological results are more mixed: some studies report that parents have lower levels of subjective well-being (Kahneman et al. 2004), while others report that fathers enjoy a higher level of life satisfaction but mothers do not (Nelson et al. 2013).

At best, we have little or no evidence that Lucky Parents is much larger than Unlucky Parents, or that Unlucky Child-frees is much larger than Lucky Child-frees. At worst, the evidence suggests that choosing to have a child is likely to reduce your overall well-being. If you reject the empirical results (which are mixed and admittedly controversial), you find yourself without evidence to guide your decision. If you accept what the balance of evidence seems to show, then the rational choice requires you to act as though your own feelings don't matter. Independently of your own feelings on the issue, you must remain childless, for those who remain childless are more likely to end up in a class of individuals that have maximized their overall utility.

Thus far, it looks like, if you accept the new decision procedure, you should either hold off on deciding, due to lack of conclusive evidence, or you should ignore your own feelings and decide to remain childless.[32] This is an interesting result. But it is *strange*. First of all, it does not bode well for the future of the species. Second, deciding *solely* on the chance that you'll end up in a class of individuals who maximized their overall utility cuts hard against the way we ordinarily consider the decision.

Imagine Sally, who has always thought that having a child would bring her happiness, deciding not to have a child simply because she knows not having one will maximize her utility. For her to choose this way, ignoring her subjective preferences and relying solely on external reasons, seems bizarre. How could Sally's own phenomenal preferences not matter to her decision? Even Lisa, who, antecedently, does not want a child, and then decides not to have a child based solely on the evidence, is not choosing in an ordinary way. Her choice, if rational,

has nothing to do with her phenomenal preferences to not have a child. Lisa does not have special insight into how she has always known that she'd be worse off as a parent: instead, she merely gets lucky. It just so happens that her phenomenal preferences support the same choice as the evidence does. Alternatively, imagine that the sizes of the classes were reversed so that Lucky Parents was much larger than Unlucky Parents, and Unlucky Child-frees was much larger than Lucky Child-frees. Now consider Anne, who has always thought that having a child would bring her misery, deciding to have a child simply because she knows it will maximize her utility. Again, the decision procedure seems bizarre from our ordinary perspective. Choosing rationally requires a very different way of thinking about the decision than we ordinarily think it does—to be rational, *we have to ignore our phenomenal preferences*.[33]

You might think that none of this applies to you. For you are a sophisticated thinker—you know, or at least you have educated, sophisticated beliefs—about which psychological characteristics really matter when you become a parent. You, unlike the unwashed masses, can judge for yourself whether you are more or less likely to end up in Lucky Parents if you have a child. I see no rational basis for a belief in such super-empirical abilities. There just isn't enough evidence available to support this sort of reasoning. Moreover, assessments of subjective well-being using the sorts of sophisticated psychological classifications that individuals would need to use to make an individually tailored, evidence-based decision are in their infancy (Kahneman and Kreuger 2006). Future empirical research might uncover the properties an individual needs to have in order to end up classed in Lucky Parents.[34] But we lack such evidence right now.[35]

As a result, the prospective parent finds herself in an interesting dilemma: ignore what she personally thinks about whether she wants to have a child and decide rationally, or take into account her own beliefs and projections about what it would be like and fail to decide rationally. Neither horn is attractive.

7. Conclusion

Contrary to popular opinion and common sense, contrary to what your parents might tell you, and contrary to the picturesque ideal romanticized by many a chick-lit novel, popular parenting guide, life coach website, and fashion magazine, you cannot rationally choose to have a child based on what you think it will be like to have a child. And, contrary to what those who are committed exclusively to their careers, or who dislike being around the children of other people, or who value their lazy weekends might believe, you cannot rationally choose to remain childless based on what you think it would have been like to have a child.

You can change the method of choosing so as to make it rational by making your choice based on something other than your phenomenal preferences. And indeed, in the past, non-subjective facts and circumstances played a much larger role in the causal process leading up to parenthood. Before contraceptive devices were widely available, you didn't choose to have a child based on what you thought it would be like. Often, you just ended up having a child. And to the extent you actively tried to choose to have children, often it was because you needed an heir, or needed more hands to work the farm, or whatever. But this is not the approach we ordinarily take now.[36] If you dispense with your phenomenal preferences, you reject a central tenet of the ordinary, twenty-first century way of thinking about the choice.

How could common sense have gotten things so wrong? I suspect that the popular conception of how to decide to have a child stems from a contemporary ideal of personal psychological development through choice. That is, a modern conception of self-realization involves the notion that one achieves a kind of maximal self-fulfillment through making reflective, rational choices about the sort of person one wants to be. (The rhetoric of the debate over abortion and medical advances in contraceptive technology have probably also contributed to the framing of the decision to have a child as a personal choice.) While the notions of personal fulfillment and self-realization through reflective choice might be apt for whether one chooses to grow one's own vegetables, what music one listens to or whether one does yoga, it is not apt for the choice to have a child. Some will conclude from my argument that we should base the decision to have a child on the values we assign to nonphenomenal outcomes or that moral considerations need to play a larger role. These conclusions might be warranted.

My view is not that it is right or wrong to have children, nor that you should not be happy with your choice, whatever choice you make.

My view is simply that you need to be honest with yourself about the basis for this choice. For example, when surprising results surface about the negative satisfaction that many parents get from having children, telling yourself that you *knew* you would not be among that class of parents, and that's why you chose to have a child, is simply a rationalization—in the wrong sense—of your act. Likewise, telling yourself that you *knew* you wouldn't be happier as a parent, and that's why you chose not to have a child, is simply an act of self-deception. You can be happy that you have a child, or happy that you are childless, without wrapping that happiness in a cloak of false rationalization.

My argument also has consequences for those who want to be able to physically conceive, carry and give birth to a child, but are unable to do so. If you want to have a child because you think having a child will maximize the values of your personal phenomenological preferences, and as a result of your inability to have a child (and thus your inability to satisfy these preferences) you experience deep sadness, depression, or other negative emotions, my argument implies that your response is not rational. This is disturbing and some might find it offensive, but it is true. Such a response is not rational. That does not mean your response is wrong, or blameworthy, or subjectively unreasonable.

All of this raises larger issues, for the sort of subjective information that experience brings is central to many of our most important personal decisions.[37] Any epistemically transformative experience that changes the self enough to generate a deep phenomenological transformation creates significant trouble for the hope that we could use our ordinary subjective perspective to make rational decisions about major life events.

ACKNOWLEDGEMENTS

I have been helped by discussion with many people. I owe special thanks to Lara Buchak, Kieran Healy, and Matt Kotzen. Thanks are also due to Erik Angner, Peter Baumann, Rachael Briggs, Ruth Chang, Tyler Doggett, Kenny Easwaran, Jordan S. Ellenberg, Daniel Gilbert, Elizabeth Harman, Hud Hudson, Jonathan Jacobs, Tania Lombrozo, Ram Neta, Alvin Plantinga, Michael Rea, Geoffrey Sayre-McCord, Tamar Schapiro, Christina van Dyke, Paul Weirich, and Mary Beth Willard.

REFERENCES

Caplan, Bryan. *Selfish Reasons to Have More Kids* (New York: Basic Books, 2011).

Evenson, Ranae, and Simon, Robin, "Clarifying the Relationship Between Parenthood and Depression," *Journal of Health and Social Behavior* 46, no. 4 (2005): pp. 341–358; http://dx.doi.org/10.1177/002214650504600403.

Hansson, S. O., "Decision Theory: A Brief Introduction," unpublished manuscript.

Jackson, Frank, "What Mary Didn't Know," *The Journal of Philosophy* 83, no. 5 (1986): pp. 291–95; https://doi.org/10.2307/2026143.

Joyce, James. *The Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press, 1999).

Kahneman, Daniel, and Kreuger, Alan, "Developments in the Measurement of Subjective Well-Being," *Journal of Economic Perspectives* 20, no. 1 (2006): pp. 3–24; http://dx.doi.org/10.1257/089533006776526030.

Kahneman, Daniel, Alan Krueger, David Schkade, Norbert Schwarz, and Arthur Stone, "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method," *Science* 306, no. 5702 (2004): pp. 1776–80; http://dx.doi.org/10.1126/science.1103572.

Levi, Isaac. *Hard Choices: Decision Making under Unresolved Conflict* (Cambridge: Cambridge University Press, 1986).

Lewis, David, "What Experience Teaches," in *Mind and Cognition: A Reader*, William Lycan, ed., (Oxford: Blackwell, 1990), pp. 499–519.

McClanahan, Sara and Adams, Julia, "The Effects of Children on Adults' Psychological Wellbeing," *Social Forces* 68, no. 1 (1989): pp. 124–46.

Nelson, S. Katherine, Kostadin Kushlev, Tammy English, Elizabeth W. Dunn, and Sonja Lyubomirsky. "In Defense of Parenthood: Children Are Associated with More Joy Than Misery." *Psychological Science* 24, no. 1 (2013): pp. 3–10.

Nomaguchi, Kei, and Milkie, Melissa, "Costs and Rewards of Children: The Effects of Becoming a Parent on Adults' Lives," *Journal of Marriage and the Family* 65, no. 2 (2003): pp. 356–74; http://dx.doi.org/10.1111/j.1741-3737.2003.00356.x.

Paul, L. A. *Transformative Experience* (Oxford: Oxford University Press, 2014).

Simon, Robin, "Life's Greatest Joy? The Negative Emotional Effects of Children on Adults," *Contexts* 7 (2008): pp. 40–45.

Ullmann-Margalit, Edna, "Big Decisions: Opting, Converting, Drifting," in *Royal Institute of Philosophy Supplements*, vol. 58, *Political Philosophy*, Anthony O'Hear, ed., (Cambridge: Cambridge University Press, 2006), pp. 157–72; http://doi.org/10.1017/S1358246106058085.

Weirich, Paul. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances* (Oxford: Oxford University Press, 2004).

Zelizer, Viviana. *Pricing the Priceless Child: The Changing Social Value of Children* (Princeton, NJ: Basic Books, 1985).

NOTES

1. My point has larger consequences for how we plan our futures and attempt to become the kind of person we think we want to be. I develop the discussion and show how my argument applies to a wide range of decisions and life experiences in Paul 2014.
2. In this example, I am assuming that you and your partner are physically able to have a child. Below, I will consider an implication of my argument for those who cannot physically produce a child. For simplicity, I am not discussing the decision to adopt a child, although I believe that a version of my argument would apply.
3. This may or may not be the same as increasing one's 'life satisfaction' or 'meaningfulness.' I will return to this at the end of the paper.
4. I am ignoring external, nonphenomenal factors one might weigh when making a choice about whether to procreate, such as the values of environmental impact or population control. A version of my argument that takes these factors into account holds unless these values are supposed to swamp the personal phenomenal values.
5. Sixty seconds of googling will turn up plenty of examples. Claims like 'You long to nurture and raise a little person who will likely be similar to you but still completely unique. Perhaps, you and your spouse feel like something is still missing, and a baby would complete your vision of family' (http://newlyweds.about.com/od/havingababy/tp/Reasons-to-Have-Kids.htm). Or see Caplan 2011. A different kind of example is provided by initiatives that try to convince young teens that they are not ready to become parents by giving them baby dolls to care for that need constant attention, wake up three times a night, etc.
6. The importance of this sort of reflective approach is underscored by the general cultural prescription against unplanned pregnancies and in the attention given to family planning by many social and religious organizations.
7. For simplicity, I am assuming a 'realist' interpretation of decision theory according to which the utility of outcomes corresponds to a real psychological quantity, such as the individual's strength of preference for outcomes or her perception of how good each outcome is. (I am indebted to Lara Buchak here.)
8. Not just anything goes. After all, the madman in the asylum can reason in accordance with his mad beliefs and come to the 'right' decision given the beliefs he started with. But his decision to follow the voices in his head and attack his fellow inmates does not conform to what we would ordinarily describe as rational behavior. The madness of his starting point—his mad beliefs—and hence the mad values he assigns to the outcomes of his choices, violate our ordinary standard. As Weirich 2004, p. 21, points out, 'an agent who maximizes utility may fall seriously short of other standards of rational action. For instance, an agent's utility assignment may be mistaken. Then, he may act irrationally even though he maximizes utility.' We can allow that an agent may rationally make a merely approximately correct utility assignment and thus act approximately rationally. The point is that the madman's original utility assignments are not rationally acceptable.
9. See Lewis 1990 for relevant discussion.
10. In other words, we are not concerned here with the debate over physicalism that the example was originally designed for.
11. In Jackson's thought experiment, because Mary has all the scientific information we'd have at the end of scientific enquiry, Mary might know what brain states will be caused by seeing red, and thus might, at least arguably, know what beliefs and desires, etc. will be caused. This kind of epistemic access is unavailable to ordinary humans reflecting on what they should do, so we can dispense with this possibility.
12. Even having a perfect duplicate of yourself around to undergo it and then tell you about the experience probably wouldn't be enough for you to know what it is like—just like a perfect duplicate couldn't tell you enough for you to know what it was like to see color if you'd never seen color before.
13. See Ullmann-Margalit 2006.
14. The phenomenological character of having a child for a blind or otherwise differently abled person will be different but just as unique.
15. Even the parent who reacts with numb disbelief or shock upon the presentation of her child has an experience with a uniquely new phenomenal character, despite the fact that the experience does not have the phenomenal character it is 'supposed' to have. Indeed, this shocked reaction could have its distinctive character in part *because* it does not have the joyous character the agent was expecting.
16. If the outcome is assigned a value based on an unacceptable mistake, the case is parallel to other cases involving decisions based on mistaken or unacceptable beliefs. '[T]ake a case in which a decision to travel by train rests on an irrational belief that the plane will crash. The decision is irrational even if it follows by utility maximization from the agent's beliefs and desires' (Weirich 2004, p. 106). Mary might believe she can assign a value to her future phenomenal state of seeing red, but she is necessarily wrong—and so if she assigns it a value, she is making an unacceptable mistake. Her belief is not rational: the value cannot be known and so her belief about it cannot be based on evidence.
17. I have been focusing on our inability to assess states with phenomenal characters that directly involve what it's like to have a child. But there are familiar knock-on effects that are less direct. Once you have a child, will you care less about your career? Will you value your child's welfare over your own? Will you still love your cat just as much? Will you love your partner more? Will you love your partner less?—Who knows? It depends on what it's like for you to have your child.

18. I'm indebted to Tamar Schapiro for this point.
19. I suppose it is one of the very few ways in which tasting Vegemite is, in fact, similar to having a child.
20. Please do not confuse this first claim with a second, different claim that adopting a child is somehow less valuable than having a child of one's own. I endorse the first claim and categorically reject the second.
21. Moreover, the claim that having a child is epistemically transformative does not entail that it is also personally transformative: for most people, it is. For some people, it isn't. But because it is epistemically transformative, you can't know whether you will find the experience personally transformative until you experience it, and so the problem for rational decision-making remains.
22. See, for example, Levi 1986 and Weirich 2004. Joyce 1999 and Hansson unpublished manuscript give excellent general discussions.
23. Weirich 2004 discusses a range of ways for agents to make normatively rational decisions under ignorance, including models where the standard for rationality is much more tolerant of ignorance. Such models permit cases that lack precise utility assignments. However, in the case of having a child, we are unable to rationally restrict the range of utilities and their probabilities in any reasonable way, preventing us from meeting even this more tolerant standard.
24. Some people find Vegemite absolutely disgusting. Others think it is delicious.
25. As Weirich points out: 'It would be difficult, even for a perfect mind, to sensibly assign intrinsic utilities to states of affairs in the absence of relevant experience. For instance, it would be difficult to assign intrinsic utility to tasting pineapple in ignorance of its taste, or to assign intrinsic utilities to eating items on the menu in an Ethiopian restaurant, even given their full descriptions, in the absence of experience with Ethiopian cuisine' (2004, p. 65).
26. I'm indebted to Elizabeth Harman for raising this objection.
27. One way to put it is to say that you need to be able to grasp the phenomenal content of the proposition described by 'Vegemite tastes disgusting,' and you can't grasp this content until you've actually tasted Vegemite. Weirich puts the point this way: 'the experience may be needed to entertain a proposition in the vivid way required for its intrinsic utility assessment' (2004, p. 66).
28. Of course, swamping can work in the other direction as well. There may be cases where the stakes are relatively low, and values of, say, certain nonphenomenal outcomes will clearly swamp the values, whatever they might be, of the relevant phenomenal outcomes. For example, if in the interest of promoting Australian tourism, foreigners receive a large financial reward for trying Vegemite for the first time, then if you are not Australian, you might rationally choose to try it on this basis. But in high stakes cases like that of having a child, one would have to make the case that such nonphenomenal outcomes exist. What is much more likely is that the value of what it is like to have the child will swamp the other outcomes.
29. I'm indebted to Kenny Easwaran for this observation.
30. McClanahan and Adams 1989 describe how a number of studies 'suggest that parenthood has negative consequences for the psychological well-being of adults.' The negative impact of children on happiness and life satisfaction has been widely discussed in sociology, psychology and economics. See, for example, Nomaguchi and Milkie 2003 and see Simon 2008 for a nice overall summary.
31. The research does show that marital status, education and financial status influence the degree to which parenthood negatively impacts emotional well-being. See Kahneman et al. 2004 and Nelson et al. 2013.
32. Depending on the context, this may amount to the same thing.
33. A way of putting the problem is like this: decision-theoretic models are constructed as tools for evaluating decisions from the third-person perspective. But our ordinary way of making personal decisions relies on the first-person perspective. This can result in a fundamental conflict.
34. Another interesting possibility is that, just by having a child, one's preferences may change in a way that changes her assessment of the value of having a child. This is directly related to the way that the experience of having a child can be both epistemically and personally transformative. If the preferences had by the prospective parent before she has a child were unchanged by the experience, they might entail that the phenomenal outcome of having a child would have a negative value. But perhaps the very fact of having the child changes the prospective parent's preferences such that the phenomenal outcome of having a child turns out to have a positive value. (There is sociological evidence that this actually happens.) This possibility raises interesting questions about how one might employ higher-order decision-theoretic structure. (I'm indebted to Tania Lombrozo here.) Ullmann-Margalit 2006 discusses related issues.
35. Frankly, I suspect that more evidence will only go so far, because the ability to determine which class one would be located in after the decision still requires a kind of self-knowledge that we can't have with epistemically transformative experiences. But that issue is beyond the scope of this discussion.
36. See Zelizer 1985 for the classic account of how children have come to be regarded as emotionally priceless.
37. I discuss this in more detail in my *Transformative Experience*, where I consider ways in which my argument applies to choices that change our phenomenological capacities, such as getting cochlear implants, and life-course-decisions such as choosing a career.

Analogy

Bertrand Russell

The postulates hitherto considered have been such as are required for knowledge of the physical world. Broadly speaking, they have led us to admit a certain degree of knowledge as to the space-time structure of the physical world, while leaving us completely agnostic as regards its qualitative character. But where other human beings are concerned, we feel that we know more than this; we are convinced that other people have thoughts and feelings that are qualitatively fairly similar to our own. We are not content to think that we know only the space–time structure of our friends' minds, or their capacity for initiating causal chains that end in sensations of our own. A philosopher might pretend to think that he knew only this, but let him get cross with his wife and you will see that he does not regard her as a mere spatio-temporal edifice of which he knows the logical properties but not a glimmer of the intrinsic character. We are therefore justified in inferring that his skepticism is professional rather than sincere.

The problem with which we are concerned is the following. We observe in ourselves such occurrences as remembering, reasoning, feeling pleasure, and feeling pain. We think that sticks and stones do not have these experiences, but that other people do. Most of us have no doubt that the higher animals feel pleasure and pain, though I was once assured by a fisherman that 'Fish have no sense nor feeling.' I failed to find out how he had acquired this knowledge. Most people would disagree with him, but would be doubtful about oysters and starfish. However this may be, common sense admits an increasing doubtfulness as we descend in the animal kingdom, but as regards human beings it admits no doubt.

It is clear that belief in the minds of others requires some postulate that is not required in physics, since physics can be content with a knowledge of structure. My present purpose is to suggest what this further postulate may be.

It is clear that we must appeal to something that may be vaguely called 'analogy.' The behavior of other people is in many ways analogous to our own, and we suppose that it must have analogous causes. What people say is what we should say if we had certain thoughts, and so we infer that they probably have these thoughts. They give us information which we can sometimes subsequently verify. They behave in ways in which we behave when we are pleased (or displeased) in circumstances in which we should be pleased (or displeased). We may talk over with a friend some incident which we have both experienced, and find that his reminiscences dovetail with our own; this is particularly convincing when he remembers something that we have forgotten but that he recalls to our thoughts. Or again: you set your boy a problem in arithmetic, and with luck he gets the right answer; this persuades you that he is capable of arithmetical reasoning. There are, in short, very many ways in which my responses to stimuli differ from those of 'dead' matter, and in all these ways other people resemble me. As it is clear to me that the causal laws governing my behavior have to do with 'thoughts,' it is natural to infer that the same is true of the analogous behavior of my friends.

The inference with which we are at present concerned is not merely that which takes us beyond solipsism, by maintaining that sensations have causes about which *something* can be known. This kind of inference, which suffices for physics, has already been considered. We are concerned now with a much more specific kind of inference, the kind that is involved in our knowledge of the thoughts and feelings of others—assuming that we have such knowledge. It is of course obvious that such knowledge is more or less doubtful. There is not only the general argument that we may be dreaming; there is also the possibility of ingenious automata. There are calculating machines that do sums much better than our schoolboy sons; there are gramophone records that remember impeccably what So-and-so said on such-and-such an occasion; there are people in the cinema who, though copies of real people,

From *Human Knowledge: Its Scope and Limits* (George Allen and Unwin, 1948), pp. 482–86. Reprinted with permission of the publisher.

are not themselves alive. There is no theoretical limit to what ingenuity could achieve in the way of producing the illusion of life where in fact life is absent.

But, you will say, in all such cases it was the thoughts of human beings that produced the ingenious mechanism. Yes, but how do you know this? And how do you know that the gramophone does *not* 'think'?

There is, in the first place, a difference in the causal laws of observable behavior. If I say to a student, 'Write me a paper on Descartes' reasons for believing in the existence of matter,' I shall, if he is industrious, cause a certain response. A gramophone record might be so constructed as to respond to this stimulus, perhaps better than the student, but if so it would be incapable of telling me anything about any other philosopher, even if I threatened to refuse to give it a degree. One of the most notable peculiarities of human behavior is change of response to a given stimulus. An ingenious person could construct an automation which would always laugh at his jokes, however often it heard them; but a human being, after laughing a few times, will yawn, and end by saying, 'How I laughed the first time I heard that joke.'

But the difference in observable behavior between living and dead matter does not suffice to prove that there are 'thoughts' connected with living bodies other than my own. It is probably possible theoretically to account for the behavior of living bodies by purely physical causal laws, and it is probably impossible to refute materialism by external observation alone. If we are to believe that there are thoughts and feelings other than our own, that must be in virtue of some inference in which our own thoughts and feelings are relevant, and such an inference must go beyond what is needed in physics.

I am, of course, not discussing the history of how we come to believe in other minds. We find ourselves believing in them when we first begin to reflect; the thought that Mother may be angry or pleased is one which rises in early infancy. What I am discussing is the possibility of a postulate which shall establish a rational connection between this belief and data, e.g., between the belief 'Mother is angry' and the hearing of a loud voice.

The abstract schema seems to be as follows. We know, from observation of ourselves, a causal law of the form 'A causes B,' where A is a 'thought' and B a physical occurrence. We sometimes observe a B when we cannot observe any A; we then infer an unobserved A. For example: I know that when I say, 'I'm thirsty,' I say so, usually, because I am thirsty, and therefore, when I hear the sentence 'I'm thirsty' at a time when I am not thirsty, I assume that someone else is thirsty. I assume this the more readily if I see before me a hot, drooping body which goes on to say, 'I have walked twenty desert miles in this heat with never a drop to drink.' It is evident that my confidence in the 'inference' is increased by increased complexity in the datum and also by increased certainty of the causal law derived from subjective observation, provided the causal law is such as to account for the complexities of the datum.

It is clear that in so far as plurality of causes is to be suspected, the kind of inference we have been considering is not valid. We are supposed to know 'A causes B,' and also to know that B has occurred; if this is to justify us in inferring A, we must know that *only* A causes B. Or, if we are content to infer that A is probable, it will suffice if we can know that in most cases it is A that causes B. If you hear thunder without having seen lightning, you confidently infer that there was lightning, because you are convinced that the sort of noise you heard is seldom caused by anything except lightning. As this example shows, our principle is not only employed to establish the existence of other minds but is habitually assumed, though in a less concrete form, in physics. I say 'a less concrete form' because unseen lightning is only abstractly similar to seen lightning, whereas we suppose the similarity of other minds to our own to be by no means purely abstract.

Complexity in the observed behavior of another person, when this can all be accounted for by a simple cause such as thirst, increases the probability of the inference by diminishing the probability of some other cause. I think that in ideally favorable circumstances the argument would be formally as follows:

From subjective observation I know that A, which is a thought or feeling, causes B, which is a bodily act, e.g., a statement. I know also that, whenever B is an act of my own body, A is its cause. I now observe an act of the kind B in a body not my own, and I am having no thought or feeling of the kind A. But I still believe, on the basis of self-observation, that only A can cause B; I therefore infer that there was an A which caused B, though it was not an A that I could observe. On this ground I infer that other people's bodies are associated with minds, which resemble mine in proportion as their bodily behavior resembles my own.

In practice, the exactness and certainty of the above statement must be softened. We cannot be sure that, in our subjective experience, A is the only cause of B. And even if A is the only cause of B in our experience, how can we know that this holds outside our experience? It is not necessary that we should know this with any certainty; it is enough if it is highly probable. It is the assumption of probability in such cases that is our postulate. The postulate may therefore be stated as follows:

> If, whenever we can observe whether A and B are present or absent, we find that every case of B has an A as a causal antecedent, then it is probable that most B's have A's as causal antecedents, even in cases where observation does not enable us to know whether A is present or not.

This postulate, if accepted, justifies the inference to other minds, as well as many other inferences that are made unreflectingly by common sense.

Intuitions About Consciousness: Experimental Studies

Joshua Knobe and Jesse Prinz

Philosophers have long been concerned with intuitions about consciousness, but this interest usually takes a peculiar form. The fundamental goal is typically not to understand the intuitions themselves, with all the psychological intricacies. Instead, what philosophers really want to understand is the true nature of consciousness, and they turn to intuitions as a way of getting indirect evidence about this other topic.

This emphasis strikes us as unfortunate. Intuitions about consciousness are fascinating phenomena, amply worthy of study in their own right. The fact that people have the intuitions they do can teach us something valuable about the way people ascribe mental states, the way they think about nonhuman animals, perhaps even the way they make moral judgments.

Our aim here, then, is to conduct a straightforward investigation into people's intuitions about consciousness. In pursuing this line of inquiry, we truly have no ulterior motives. It is not as though we are trying to present a theory about the true nature of consciousness and have simply chosen to argue for it in a roundabout way. Rather, we are genuinely intrigued by the intuitions themselves, and we want to get a better understanding of the psychological mechanisms that generate them. Our paper therefore draws on a number of different lines of existing research, including research in 'theory of mind' (e.g., Gopnik and Meltzoff 1997; Scholl and Leslie 1999), research in consciousness studies (e.g., Block 1978, 1995), and research about how people determine which sorts of entities are capable of having mental states (Inagaki and Hatano 1991; Johnson 2000).

Because our aims are somewhat unusual, we will be making use of a somewhat unusual method. First we introduce hypotheses about the psychological mechanisms underlying people's intuitions; then we put these hypotheses to the test using systematic experiments.

I.

We begin by setting out two initial hypotheses. These hypotheses will not be concerned directly with the actual patterns of people's intuitions. Instead, they will be concerned with certain underlying psychological processes. But when the two hypotheses are put together (and combined with a few plausible assumptions), they yield definite testable predictions.

1. Our first hypothesis is that ordinary people—people who have never studied philosophy or cognitive science—actually have a concept of phenomenal consciousness. In particular, we hypothesize that

From *Phenomenology and the Cognitive Sciences* 7: 67–83, 2007. Reprinted by permission of the publisher.

people often make use of the concept of phenomenal consciousness when they are ascribing mental states. Thus, suppose that a person is wondering whether or not to make the ascription:

(1) Sasha is now experiencing great pain.

The person will recognize that mental state ascriptions like (1) require phenomenal consciousness. Hence, she will ascribe the state only if she believes that the agent under discussion (in this case, a person named 'Sasha') is capable of having phenomenally conscious states.

To a first glance, this first hypothesis may seem a bit absurd. After all, it is clear that most people would not understand the *words* 'phenomenal consciousness,' and when one tries to explain the concept in a classroom, students often have trouble understanding what it amounts to. It would certainly be foolish, then, for us to suggest that people ordinarily have explicit beliefs about whether particular mental state types do or do not require phenomenal consciousness. But that is not at all what we have in mind. What we mean to suggest is rather that people grasp this concept at a purely tacit level. In other words, the suggestion is that people are actually applying the concept all the time; it's just that they normally have no awareness of doing so.

To a get a sense for what we mean here, it might be helpful to consider the way research typically proceeds in linguistics. Linguists quite often suggest that people categorize words or phrases using a particular concept even when they have no awareness at all of doing so. For a simple example, consider the following four phrases:

(2) a. professorial gentleman
b. regal style
c. financial planner
d. criminal investigation

A linguist might say that we are actually categorizing the adjectives in these phrases using various complex concepts—e.g., that we are classifying the adjectives in the first two phrases (2a–b) as 'qualitative adjectives' and those in the latter two phrases (2c–d) as 'relational adjectives.' Of course, we are not normally aware of making any such distinction, but one can see that some such thing must be going on when one tries to change the phrases around as follows:

(3) a. The gentleman was professorial.
b. The style was regal.
*c. The planner was financial.
*d. The investigation was criminal.

What we see here is that people find it acceptable to use the adjectives *professorial* and *regal* in contexts where they don't find it acceptable to use *financial* or *mayoral*. Evidence like this can offer us insight into the ways in which people ordinarily classify these words.

Our aim here is to use a similar strategy to study ascriptions of mental states. Thus, consider the following four mental state ascriptions:

(4) a. Sasha is vividly imagining a purple square.
b. Sasha is experiencing intense joy.
c. Sasha is wondering what to do.
d. Sasha is considering his options.

Our hypothesis is that people have a concept of phenomenal consciousness and that they use this concept to distinguish between different types of mental state ascriptions. Specifically, we would suggest that people are classifying the first two ascriptions (4a–b) as ascriptions that require phenomenal consciousness and that they are classifying the latter two (4c–d) as ascriptions that do not require phenomenal consciousness. We propose to provide evidence for that claim by showing that it is possible to change around all four ascriptions in the same way, such that people will find the revised versions of (4c–d) acceptable, but they will regard the revised versions of (4a–b) as completely incorrect.

2. Before we can propose our second hypothesis, we need to introduce a few technical terms:
   - The *functional role* of a state is the profile of its typical causes and effects. If we wanted to characterize the functional role of anger, we might mention that people often get angry when they perceive themselves to be victims of a slight, that angry people often seek some form of revenge, and so on.
   - The *physical constitution* of an entity is its actual physical make-up. If we wanted to characterize the physical constitution of a human being, we might say that human beings have four limbs and a head, that this head contains a brain, that the brain contains neurons, and so forth.

The key point for present purposes is that, even when two entities are extremely different in their physical constitutions, there can sometimes be a certain sort of isomorphism between their states. Specifically, it can sometimes be possible to map the states of the first entity onto

those of the second in such a way that all of the causal generalizations that apply to types of states in the first entity also end up applying to types of states in the second. In such cases, we will say that the states of the two entities have *similar functional roles*.[1]

Now, suppose we encountered an entity whose physical constitution was very different from our own but whose states were extremely similar to our mental states in their functional roles. It would be possible to predict and explain this entity's behaviors using all of the same causal generalizations we normally apply to ourselves. We could ascribe to the entity states of belief, intention, anger, etc. and then use these ascriptions to generate predictions that would be just as accurate as those we would make of an ordinary human being. Still, it seems that an important question remains. We might know that we could predict the entity's behavior extremely well by *saying* that it was angry. . . but there remains an open question as to whether or not the entity truly would *be* angry.

This question has been discussed at great length in recent work in the philosophy of mind. In trying to answer it, philosophers have drawn on arguments from metaphysics, philosophy of science, logic and, above all, people's intuitions about particular cases.[2] But a funny thing has happened as research in this tradition proceeds. Although philosophers first became interested in people's intuitions because they wanted to solve a more general problem in the philosophy of mind, it has gradually become clear that people's intuitions show surprisingly intricate patterns and are worthy of study in and of themselves. It is this sort of study that we take up here.

We come then to our second hypothesis. This hypothesis is that information about physical constitution plays different roles in different kinds of mental state ascriptions. Specifically, information about physical constitution plays a special role in those ascriptions that require phenomenal consciousness—a role that it does not play in other kinds of mental state ascription.

Before going any further, we need to get clear about what this hypothesis does and does not entail. It certainly does not say anything very specific either about the process underlying ascriptions of states requiring phenomenal consciousness or about ascriptions of other sorts of states. All it says is that there is a particular type of *difference* between these different kinds of ascriptions. Still, the claim it makes is a fairly surprising one. One might have thought that there was some general truth about how people ascribe mental states and that, whatever people turned out to be doing with information about physical constitution, they would at least do the very same thing for all kinds of mental state ascriptions. Our second hypothesis denies this. It asserts that ascriptions of states that require phenomenal consciousness are governed by certain special criteria that do not apply to any other mental state ascriptions.

3. Putting these two hypotheses together, we arrive at a new prediction. From the first hypothesis we learn that people classify mental states like those ascribed in (4a) as requiring phenomenal consciousness.

   (4) a. Sasha is vividly imagining a purple square.

and that they regard mental states like those ascribed in (4)c as not requiring phenomenal consciousness.

   (4)c. Sasha is wondering what to do.

But from the second hypothesis, we learn that the process underlying ascriptions of states that require phenomenal consciousness makes use of information about physical constitution in a way that other mental ascriptions do not. We now arrive at a somewhat surprising prediction. Suppose that we switch around both of these sentences by getting rid of the word 'Sasha' in both and replacing it with a description of a very different sort of entity. This entity would be capable of having states with functional roles that resembled those of human mental states, but it would be radically different from a human being from a physical perspective. If we chose just the right sort of entity, we should find that people regard ascriptions to it of states that do not require phenomenal consciousness as perfectly acceptable but that they should regard ascriptions to it of states that do require phenomenal consciousness as completely wrong.

II.

In thinking about these issues, philosophers often resort to bizarre science-fictional entities like giant computers made of strung-together soda cans or robots controlled by troops of miniature girl scouts. Our focus here will be on examples of a more pedestrian variety. We will be concerned with entities like corporations, clubs and nations. In other words, we will be concerned with the sorts of entities usually referred to as *group agents*.

From the standpoint of physical constitution, group agents are radically different from individual human beings. In individual humans, decision-making is realized by neurons, synapses and firing rates. In a group agent, decision-making might be realized by committees, memos and emails. Clearly, the decision making of group agents can be realized by physical objects that have no parallel in individual humans.

And yet, we do often ascribe mental states to group agents. It seems quite natural to say that Microsoft 'intends' to release a new product or that it 'believes' that Netscape is one of its main competitors. Presumably, our willingness to ascribe these mental states stems not from a similarity in physical constitution but from a similarity in functional roles.

In saying all this about group agents, we are simply echoing the view that has become standard among researchers in the field. But someone might object to that view. He or she might say:

> When people say that a corporation 'intends' to do something, they aren't really ascribing a mental to a group as such. What they mean is that certain *members* of the group have the mental state in question. Sometimes they have a funny way of expressing themselves, but that is just some kind of metaphor or shorthand.

Most researchers who have thought seriously about these questions would reject objections like this one. They think that the expressions under discussion here are not just shorthand and that people really are ascribing mental states to groups. (For discussion, see Bloom and Veres 1999; Gilbert 1992; Huebner 2008; Kashima et al. 2005; Morris et al. 2001; O'Laughlin and Malle 2002; Pettit 2003; Searle 1995; Solan 2005; Tollefsen 2002; Tuomela 1995; Velleman 1997.)

Before we go on to describe our own experiments, it might be helpful to review some of the arguments that have convinced researchers of this view.

Huebner's argument Consider the way we ordinarily go about ascribing mental states to a person. We might think that the person's behavior is ultimately determined by certain patterns of neural activity in his or her brain, but we do not typically try to explain the person's behavior in terms of the states of individual neurons. Instead, we use more abstract psychological generalizations. These generalizations can be considered 'robust' in the sense that they would continue to hold even if the properties of the individual neurons had been somewhat different.

Now consider the way we ordinarily ascribe mental states to groups. We might believe that the behavior of the group as a whole ultimately depends on the activities of the individual members, but we often explain group behavior in a way that does not make explicit reference to any specific members. Instead, we rely on more abstract generalizations about the nature of group behavior. These generalizations can also be considered 'robust,' since they would continue to hold even if the properties of the individual members changed in various ways.

In short, the relationship between ascriptions of states to a group and ascriptions of states to individual members is more or less the same as the relationship between ascriptions of states to a person and ascriptions of states to his or her individual neurons. (Huebner 2008).

Velleman's argument A philosophy department can intend to hire a new professor. But it seems that (a) one can't intend to perform a behavior if one doesn't think oneself capable of performing that behavior and (b) no individual member of the department believes him or herself capable of hiring a new professor. Therefore, the only entity that can have the intention is the department itself (Velleman 1997).

The Arizona Experimental Philosophy Lab's argument The Arizona EPL ran a study in which they explicitly asked subjects whether ascriptions of mental states to corporations were literal or figurative. Subjects were given a series of sentences and asked to rate them on a scale from 1 ('figurative') to 7 ('literal'). The key sentence was: 'Some corporations want lower taxes.' Subjects gave this sentence a rating of 6.2—a truly resounding vote against the view that such sentences are purely figurative. (Arico et al. 2006).

In what follows, we build on the work of these earlier researchers. We assume that people truly are ascribing mental states to groups and then use these ascriptions as a way of getting a handle on the structure and function of people's concept of phenomenal consciousness.

III.

It is a striking fact about group agents that we ascribe to them some types of mental states but not others. We might say that Microsoft intends something or wants something or

believes something . . . but there are other kinds of ascriptions that we would never make to Microsoft. For example, we would never say that Microsoft was feeling depressed. This is a puzzling phenomenon, and one can learn a lot about ordinary mental state ascription by trying to understand how it arises.

First of all, it should be emphasized that a state of a corporation easily could have a functional role similar to the one that people ordinarily associate with feeling depressed. So, for example, suppose that Microsoft had a department in charge of monitoring net cash flow. When cash flow becomes too low, it sends out a warning to all other departments of the corporation. Those other departments then stop moving forward on the projects they had previously been pursuing and instead take time to reflect on any mistakes they might have been making in their overall approach. This state, or something very much like it, would show the profile of causes and effects normally associated with feelings of depression. Or, at a very minimum, it would be just as similar in function role to depression as the 'intentions' of a corporation are to those of a human individual.

Yet the fact remains that people do not normally ascribe feelings of depression to group agents. We suspect that their unwillingness to make such ascriptions has nothing to do with dissimilarities in functional roles. Instead, we propose to explain it in terms of our two hypotheses:

(1) People tacitly classify *feeling depressed* as a state requiring phenomenal consciousness
(2) Ascriptions of states requiring phenomenal consciousness are sensitive in a special way to information about physical constitution.

To test this explanation, we conducted a series of experiments.

Study 1

The first thing we need to show is that people do ascribe various mental states to group agents but that they do not ascribe states that require phenomenal consciousness. Here we do not simply want to know whether people will be willing to ascribe certain states when pressed; we also want to know whether people are naturally inclined to ascribe those states in ordinary life.

The first author's beloved wife, Alina Simone, came up with the perfect solution. She entered into Google a series of phrases that ascribed mental states to group agents. Some ascribed phenomenal states; some ascribed non-phenomenal states. By comparing the number of hits that each type of ascription received, we can see whether people are more inclined to ascribe certain types of states than they are to ascribe others.

Here are the phrases ascribing non-phenomenal states, along with the number of hits that each phrase received:

'Microsoft intends'	25,700
'Microsoft decides'	11,400
'Microsoft tries'	52,600
'Microsoft wants'	135,000
'Microsoft believes'	31,100
'Microsoft hopes'	56,600
'Microsoft loves'	4,100
'Microsoft hates'	970

And here are the phrases ascribing phenomenal states:

'Microsoft feels depressed'	0
'Microsoft experiences joy'	0
'Microsoft feels happy'	0
'Microsoft feels pain'	2
'Microsoft feels angry'	0
'Microsoft feels scared'	0

The difference between the number of hits received by phrases in each of these two groups was so dramatic that, even with such a small sample size, one actually obtains a statistically significant effect.[3]

But now we face a problem. We know that people use certain English expressions more frequently than others, but we do not know precisely *why* they do this. It could be that the whole effect is due to some trivial difference like the number of words contained in each expression or the frequency with which people generally ascribe different types of states. What we want to know now is whether people are refraining from ascribing certain states to group agents because they actually regard those ascriptions as unacceptable. To address this question, we ran a series of studies on human subjects.

Study 2

We began with a study in which subjects were given a list of sentences that ascribed mental states to group agents and then asked whether

each of these sentences 'sounds natural' or 'sounds weird.' Some of the sentences ascribed non-phenomenal states; others ascribed phenomenal states. The two types of sentences were mixed together, and the order of presentation was randomized.

The sentences ascribing non-phenomenal states were:

- Acme Corp. believes that its profit margin will soon increase.
- Acme Corp. intends to release a new product this January.
- Acme Corp. wants to change its corporate image.
- Acme Corp. knows that it can never compete with GenCorp in the pharmaceuticals market.
- Acme Corp. has just decided to adopt a new marketing plan.

The sentences ascribing phenomenal states were:

- Acme Corp. is now experiencing great joy.
- Acme Corp. is getting depressed.
- Acme Corp. is feeling excruciating pain.
- Acme Corp. is experiencing a sudden urge to pursue internet advertising.
- Acme Corp. is now vividly imagining a purple square.

For each sentence, subjects were asked to provide a rating on a scale from 1 ('sounds weird') to 7 ('sounds natural'). The mean ratings were as follows:

Non-phenomenal states:

6.6: Deciding
6.6: Wanting
6.3: Intending
6.1: Believing
5.2: Knowing

Phenomenal states:

4.7: Experiencing a sudden urge[4]
3.7: Experiencing great joy
2.7: Vividly imagining
2.5: Getting depressed
2.1: Feeling excruciating pain

As the table shows, even the most acceptable phenomenal state was still deemed less acceptable than the least acceptable non-phenomenal state. More generally, there was a statistically significant effect such that people gave lower ratings for the phenomenal states than for the non-phenomenal states.[5]

Study 3

The results reported thus far seem to indicate that people are unwilling to ascribe to group agents states that require phenomenal consciousness. We now turn to questions about precisely what sort of criteria people are applying here. What exactly is it about group agents that makes people regard them as unable to have certain kinds of states?

One possibility would be that subjects' judgments are based on *similarity to humans*. Subjects start out with the premise that human beings have phenomenal consciousness. Then, when they are wondering whether some other sort of agent has phenomenal consciousness, they simply ask whether its physical constitution is sufficiently similar to that of human beings. Since the physical constitution of a corporation is extremely unlike that of a human being in numerous respects, subjects conclude that corporations do not have phenomenal consciousness.

But there is also another possibility. Perhaps subjects are not thinking at all about similarity to human beings. Perhaps they are applying a far more specific restriction on constitution (say, a restriction against agents that are composed of other agents). On this latter view, people might be willing to ascribe phenomenal states to agents that are very, very different from us—just as long as those agents do not violate the specific restriction.

To decide between these conflicting hypotheses, we ran a follow-up experiment. All subjects were given a description of an agent that is not in any sense made up of smaller agents but which nonetheless has a physical constitution radically different from our own:

> Once there was a powerful sorceress. She came upon an ordinary chair and cast a spell on it that endowed it with a mind. The chair was still just made of wood, but because of the magic spell, it could now think complex thoughts and form elaborate plans. It would make detailed requests to the people around it, and if they didn't do everything just as it wanted, it would start complaining. People used to call it the Enchanted Chair.

Note that this passage ascribes to the chair only states that do not require phenomenal consciousness. (Indeed, it only ascribes states that people would be perfectly happy to ascribe to a corporation.) The key question now is whether people will automatically conclude that the chair is also capable of having states that

require phenomenal consciousness. Subjects were therefore asked the question: 'Can the Enchanted Chair *feel happy or sad*?'

In addition, all subjects were also given a brief description of the Acme Corporation. They were then asked a question designed to see whether they would ascribe phenomenal states to that corporation, namely: 'Can Acme Corp. *feel happy or sad*?'

Both answers were given on a scale from 1 to 7. Subjects once again refused to ascribe phenomenal states to the corporation (average rating: 1.8), but they were happy to ascribe phenomenal states to the chair (average rating: 5.6). This difference was statistically significant.[6]

The moral here is clear. From the standpoint of physical constitution, a wooden chair is extremely different from a human being. Yet people were perfectly willing to ascribe phenomenal states to the chair. It therefore appears that people do not simply refuse to ascribe phenomenal states to any agent that differs from human beings in its physical constitution. They must be making use of some more specific restriction that rules out group agents on independent grounds. In philosophical jargon, our respondents were committed to the multiple realizability of phenomenal states, but they were also willing to impose certain specific restrictions on physical constitution.

Study 4

The results of these first three experiments show that people are not willing to apply certain kinds of sentences to group agents. It seems that a full explanation of this effect would consist of two basic parts. First it would provide an account of the way in which people map the actual words in the sentences onto various underlying concepts; then it would provide an account of why people are unwilling to apply these concepts to group agents.

For a simple example, consider the fact that people seem unwilling to say that a group agent can be 'feeling upset.' What we want now is a step-by-step explanation of the process that leads up to this intuition.

Here is one possible view. First people map the phrase 'feeling upset' onto the concept *upsetness*; then they determine that no group agent can satisfy the criteria associated with the concept of upsetness. This view is a plausible one, but we suspect that it is actually incorrect.

Instead, we want to propose a slightly more complex account. When people hear the phrase 'feeling upset,' they recognize that this phrase cannot correctly be applied to an agent unless that agent fulfills both the criteria associated with the concept *upsetness* and the criteria associated with the concept *phenomenal consciousness*. There is actually no obstacle to a group agent fulfilling the criteria associated with the concept of upsetness. People's reluctance to apply phrases like this one to group agents derives entirely from the criteria associated with the concept of phenomenal consciousness.

In other words, people should be perfectly willing to ascribe upsetness to a corporation. The problem is simply that they don't think corporations are capable of genuinely *feeling* anything. If they had some way of saying that a corporation was in a state of being upset without implying that the corporation actually felt anything, they should be perfectly happy to do so.

To test this hypothesis, it would be helpful to find a way of holding fixed the degree to which people ascribe upsetness and varying only the degree to which they ascribe phenomenal consciousness. Consider, in this light, the following pair of sentences:

(5) a. Acme Corp. is feeling upset.
    b. Acme Corp. is upset about the court's recent ruling.

Here it seems that both sentences ascribe upsetness to a corporation. The chief difference between them is just that only the first sentence ascribes phenomenal consciousness. The second sentence seems to indicate that the corporation is in a state of upsetness without also indicating that the corporation was genuinely capable of having feelings.

Similarly, consider the pair:

(6) a. Acme Corp. is feeling regret.
    b. Acme Corp. regrets its recent decision.

Here is seems that both sentences ascribe regret to a corporation, but only the first also ascribes phenomenal consciousness.

In our fourth study, we presented these two pairs of sentences to subjects and asked them to rate the sentences in each pair on a scale from 1 ('sounds weird') to 7 ('sounds natural'). The mean responses were as follows:

	With 'Feeling'	Without 'Feeling'
Upset	1.9	5.3
Regret	2.8	6.1

Note that subjects gave far higher ratings to the sentences that did not include the

word 'feeling.' This difference is statistically significant.

Looking at these results, it seems clear that people are not showing an across-the-board tendency to reject ascriptions of upsetness and regret to group agents. On the contrary, it seems that people are perfectly willing to say that a group agent can be in a state of upsetness or regret. The problem is simply that it cannot *feel* upset or *feel* regret. In short, it seems that people's reluctance to say that a group agent 'feels upset' stems not from the criteria associated with their concept of upsetness but rather from the criteria associated with their concept of phenomenal consciousness.

IV.

Thus far, we have been presenting data to indicate that people's ascriptions of consciousness are sensitive not only to functional information but also to information about physical constitution. We now ask whether our findings might have any implications for broader questions about the nature of folk psychology.

To begin with, we can ask why anyone would have thought that folk psychology was functionalist in the first place. Clearly, the answer is not that researchers derived this conclusion from empirical studies of mental state ascriptions. That is, it is not as though cognitive scientists just went out and studied a lot of different kinds of mental state ascriptions, found that all of them were best understood functionally, and then concluded that folk psychology as a whole was probably functionalist. Instead, it seems that the idea that folk psychology is fundamentally functionalist was derived from a far broader view—a kind of *grand vision* of the nature of folk psychology.

This grand vision says that folk psychology should be understood, most fundamentally, as a tool for predicting and explaining behavior. Researchers who subscribe to this vision often suggest that folk psychology is in many ways similar to a scientific theory. Just as a scientist might posit unobservable entities in order to predict and explain the behavior of the observables, so too the folk psychologist posits unobservable mental states as a way of predicting and explaining human behavior. The key claim here is that we will be able to understand why people ascribe mental states in precisely the way they do if we reflect on the ways in which these ascriptions facilitate the activities of prediction and explanation.

Starting from this grand vision, it is only a short step to the view that folk psychology must be functionalist. After all, if the vision is correct, it seems that the only properties of mental states that could play a role in folk psychology are those properties that might contribute to prediction and explanation—and the only properties that could be helpful in prediction and explanation are those that have something to do with the state's causes and effects. This chain of reasoning strikes us as a powerful and compelling one.

Yet the results reported here have moved us to accept a theory that does not fit well with the functionalist view. On this theory, people only ascribe phenomenal states to entities that satisfy a complex system of non-functionalist principles. It is hard to see how a psychological mechanism like this one could be best understood as a tool for predicting and explaining behavior.

To bring out the problem here, it might be helpful to emphasize that we seem to be uncovering a mechanism that specifically blocks the ascription of certain mental states even in cases where ascriptions of those mental states would facilitate prediction and explanation. Thus, suppose we find that we can do a better job of predicting and explaining the behavior of a given entity if we sometimes ascribe to it feelings of depression. If the entity in question has the wrong type of physical make-up—e.g., if it is a group agent—a special type of psychological mechanism will kick in and block the ascription of depression to the entity. How exactly could such a mechanism be understood as a tool for facilitating prediction and explanation?

If we had overwhelming evidence for the thesis that all aspects of folk psychology were best understood in terms of prediction and explanation, the right thing to do now would be to introduce some ad hoc assumption that allowed us to reconcile the thesis with our data. But the truth is that this thesis has been under attack in a number of other domains. In fact, a growing body of experimental research now points to a rather different picture of the nature of folk psychology. This research suggests that, although some aspects of folk psychology may indeed be best understood as tools for prediction and explanation, others are best understood in terms of their role in facilitating *moral* judgment (e.g., Cushman 2006; Knobe 2005, 2007; Leslie et al. 2006; McCann 2005; Mele 2003).

In other words, the suggestion is that we won't be able to make sense of every last aspect of folk psychology just by thinking about the importance of prediction and explanation. Some aspects of folk psychology will only begin to make sense when treat them as tools for facilitating judgments about what is right or wrong, praiseworthy or blameworthy.

In light of this new wave of research, we think it would be a mistake to suppose that there must be some way to understand our findings in terms of the use of folk psychology in prediction and explanation. Instead, we suggest that the best approach would be to consider *all* of the various uses to which phenomenal state ascriptions are put. Then we can ask whether the findings can actually be understood more simply or elegantly in terms of some other use.

Let us try, then, to put aside our theoretical preconceptions and take a fresh look at the phenomena. We can consider a prototypical case in which someone might wonder whether an entity is capable of having phenomenal states and then ask how this sort of question is best understood. Suppose, for example, that we are observing a fish that has been injured and is squirming about helplessly. If we had to say what mental state this fish was in, we might be tempted to say that it is in 'pain.' Yet it does seem that there is still a legitimate question as to whether or not the fish is truly capable of phenomenal consciousness. Thus, a person might well think to herself: 'I see that the fish is squirming, but is it truly capable of feeling pain? Can a fish truly *feel* anything at all?' The experiments reported above suggest that people actually engage in certain highly complex psychological processing when trying to address questions like this one. What we want to understand now is what role all of this psychological processing actually serves in their lives.

It is certainly a bit difficult to see how all of this processing could be justified in terms of its potential to facilitate future behavioral predictions. (In fact, our bet would be that this processing doesn't end up facilitating behavioral prediction at all.) Yet, it isn't at all difficult to see how the answer to this question might play a role in a person's future decisions. Regardless of whether it in any way facilitates behavioral predictions, it can certainly influence a person's subsequent *moral* judgments. The more certain we are that an entity is capable of having phenomenal states, the more certain we will be that it is important to treat that entity with moral concern.

This point comes out especially clearly in recent work by Gray et al. 2007 and Jack et al. 2006. These researchers conducted experiments to see what sorts of mental state ascriptions most affected people's moral judgments. The results showed that people's judgments that a given entity was worthy of moral concern were affected far more by ascriptions of phenomenal states than by ascriptions of other sorts of mental states. In other words, when we are wondering whether to treat an entity with moral concern, we are not principally concerned with questions about whether this entity is capable of complex reasoning, planning or comprehension—what we really want to know is whether or not the entity is capable of having genuine feelings.[7]

Now, if we focus on these moral concerns, it becomes easy to see why people might find it important to determine whether or not a given agent has phenomenal consciousness. There is no need to construct some complex story about how ascriptions of phenomenal consciousness might actually be able to facilitate behavioral prediction. Instead, we can simply rely on the straightforward idea that ascriptions of phenomenal consciousness have an impact on subsequent moral judgments. Take the person who is staring at the fish and wondering whether it is genuinely capable of phenomenal consciousness. There is a clear and rather obvious sense in which the question she is asking might be relevant to her future behaviors. She needs to know whether fish can truly feel pain because she needs to know what sorts of moral obligations she has toward the fish.

V.

In discussing this question with other researchers, we find that many of them see us as proposing a radical and counterintuitive doctrine. We think that this reaction gets things exactly backwards. While it is true that many cognitive scientists regard all aspects of folk psychology as tools for behavioral prediction, we think that *they* are the ones upholding a radical and counterintuitive doctrine. Meanwhile, we see ourselves as simply standing up for the commonsense view.

To address this issue, we ran one final experiment—this time, with the aim of getting a handle on ordinary people's views about the functions served by ascriptions of consciousness.

Study 5

Subjects were asked to frame their own hypotheses about why people might be interested in ascribing certain kinds of mental states. Some subjects were asked why people might be interested in ascribing a capacity for memory; others were asked why people might be interested in ascribing a capacity for consciousness. We thought that subjects would offer different hypotheses for these different kinds of ascriptions. Specifically, we thought that they would naturally tend to explain ascriptions of memory in terms of the aim of prediction and explanation but that they would spontaneously explain ascriptions of consciousness in terms of a need to form moral judgments.

Subjects in the *memory condition* received the following question:

> Imagine a person who has a job working with fish. He finds himself wanting to know the answer to a particular question about them. Specifically, he wants to know whether fish are capable of *remembering* which part of a lake has the most food.
>
> Why do you think he might want to know this? Why might the question be important to him?

Subjects in the consciousness condition received a question that was almost exactly the same, except that the person was described as wondering whether fish were 'genuinely capable of feeling anything':

> Imagine a person who has a job working with fish. He finds himself wanting to know the answer to a particular question about them. Specifically, he wants to know whether fish are genuinely capable of *feeling* anything.
>
> Why do you think he might want to know this? Why might the question be important to him?

After reading each question, subjects provided a free-response answer in the space below. These answers could then be coded into categories for statistical analysis.

First, we went through each of the responses and determined whether or not it said that the man would be interested in ascribing the relevant capacity for reasons having to do with *prediction, explanation or control*. Here is an example of a response that was classified in this first category:

> So it will be easier to feed them, b/c he only has to distribute food in one place or so he'll know where to go in order to give bait, if they are capable of remembering such things.

Second, we went through each response and determined whether or not it said that the man would be interested in ascribing the relevant capacity for reasons that had to do with making *moral judgments*. Here is a response that was classified in this second category:

> He might want to know whether fish genuinely feel things because in doing his job, he does lots of things to the fish that might possibly hurt them if they can really feel things. It might be important to him to find out if he causes them pain because he might feel it is unethical or immoral to cause harm to other things. He could hold this belief for several reasons such as religion.

These two kinds of categorization were performed independently, so that any given response could be coded into one category, into both, or into neither.

Overall, responses in the memory condition fit well with the traditional 'grand vision' about the function of folk psychology. Subjects overwhelmingly responded that the man would be interested in ascribing a capacity for memory because this ascription could enable him to predict, explain or control behavior. (100% of responses referred to prediction, explanation or control; 9% referred to moral judgment.)

But responses in the consciousness condition were very different. In that condition, subjects *did not* refer to an interest in prediction, explanation or control. Instead, the overwhelming tendency was to explain these ascriptions in terms of an interest in moral judgment. (0% of responses referred to prediction, explanation or control; 100% referred to moral judgment.)

Of course, it is possible that people will turn out to be mistaken here. That is, it is possible that people *believe* that they are interested in these questions primarily for moral reasons but that they are *really* interested in these questions primarily as a way of facilitating subsequent prediction, explanation and control. Yet, although this sort of mistake is possible, we see no specific reason to believe that it is taking place here. Indeed, we see no reason at all to think that ascriptions of consciousness are best understood as tools for the prediction, explanation and control of behavior.

VI.

We therefore tentatively offer a more complex account of the function of these ascriptions. Suppose, e.g., that a person concludes

(7) George is feeling upset.

The evidence provided above suggests that it would be a mistake to consider this judgment as a whole and ask what role it might serve in people's lives. Instead, we need to break it down into two parts—an ascription of upsetness and an ascription of phenomenal consciousness and consider each of them separately. It may very well turn out that ascriptions of upsetness serve primarily to facilitate behavioral prediction, but it does not appear that this same approach can be helpfully applied to ascriptions of phenomenal consciousness. On the contrary, it seems that ascriptions of phenomenal consciousness are best understood in terms of their role in facilitating moral judgment.

ACKNOWLEDGEMENT

For helpful comments and suggestions, we wish to thank Ned Block, Paul Bloom, Bryce Huebner, Tony Jack, Uriah Kriegel, Tania Lombrozo, Bill Lycan, Bertram Malle, Ram Neta, Shaun Nichols, Philip Pettit, David Velleman, and the anonymous author of Mixing Memory. We are especially grateful to Kriegel, whose incisive comments led to major changes in a number of aspects of our paper.

REFERENCES

Arico, A., B. Fiala, and S. Nichols. *The Folk Psychology of Consciousness* (unpublished manuscript, University of Arizona, 2006).

Block, N., "Troubles with Functionalism," in *Minnesota Studies in the Philosophy of Science, vol. IX*, C. Wade Savage, ed., (Minneapolis: University of Minneapolis Press, 1978), pp. 261–325.

_____ "On a Confusion About the Function of Consciousness," *Behavioral and Brain Sciences* 18 (1995): pp. 227–47.

Bloom, P., and Veres, C., "The Perceived Intentionality of Groups," *Cognition* 71 (1999): pp. B1–B9.

Cushman, F. *Judgments of Morality, Causation and Intention: Assessing the Connections* (unpublished manuscript, Harvard University, 2006).

Gilbert, M. *On Social Facts* (Princeton, NJ: Princeton University Press, 1992).

Gopnik, A., and Meltzoff, A. N. *Words, Thoughts, and Theories* (Cambridge, MA: MIT Press, 1997).

Gray, H., K. Gray, and D. Wegner, "Dimensions of Mind Perception." *Science* 315 (2007): p. 619.

Huebner, B., PhD dissertation, University of North Carolina-Chapel Hill, 2008.

Inagaki, K., and Hatano, G., "Constrained Person Analogy in Young Children's Biological Inference," *Cognitive Development* 6 (1991): pp. 219–31.

Jack, A. I., A. Roepstorff, and P. Robbins, *The Genuine Problem of Consciousness: Trusting the Subject* (2006). Manuscript under revision.

Johnson, S., "The Recognition of Mentalistic Agents in Infancy," *Trends in Cognitive Sciences* 4 (2000): pp. 22–28.

Kashima, Y., E. Kashima, C-Y. Chiu, T. Farsides, M. Gelfand, Y-Y. Hong, et al., "Culture, Essentialism, and Agency: Are Individuals Universally Believed to be More Real Entities Than Groups?" *European Journal of Social Psychology* 35 (2005): pp. 147–69.

Knobe, J., "Theory of Mind and Moral Cognition: Exploring the Connections," *Trends in Cognitive Sciences* 9 (2005): pp. 357–59.

_____"Folk Psychology: Science and Morals," in *Folk Psychology Reassessed,* D. Hutto and M.Ratcliffe, eds., (Kluwer/Springer Press, 2007).

Knobe, J., and Prinz, J. *Experimental Studies of Intuitions About Consciousness: Methodological and Statistical Details.* (2006)

Leslie, A., J. Knobe, and A, Cohen, "Acting Intentionally and the Side-Effect Effect: 'Theory of Mind' and Moral Judgment," *Psychological Science* 17 (2006): pp. 421–27.

McCann, H., "Intentional Action and Intending: Recent Empirical Studies," *Philosophical Psychology* 18 (2005): pp. 737–48.

Mele, A., "Intentional Action: Controversies, Data, and Core Hypotheses," *Philosophical Psychology* 16 (2003): pp. 325–40.

Morris, M., T. Menon, and D. Ames, "Culturally Conferred Conceptions of Agency: A Key to Social Perception of Persons, Groups, and Other Actors," *Personality and Social Psychology Review* 5 (2001): pp. 169–82.

O'Laughlin, M., and Malle, B. F., "How People Explain Actions Performed by Groups and Individuals," *Journal of Personality and Social Psychology* 82 (2002): pp. 33–48.

Pettit, P., "Groups with Minds of Their Own," in *Socializing Metaphysics,* F. Schmitt, ed.,. (New York: Rowman & Littlefield, 2003), pp. 167–93.

Searle, J. *The Construction of Social Reality* (New York: Free Press, 1995).

Scholl, B. J., and Leslie, A. M., "Modularity, Development and 'Theory of Mind.'" *Mind & Language* 14 (1999): pp. 131–53.

Solan, L., "Private Language, Public Laws: The Central Role of Legislative Intent in Statutory Interpretation," *Georgetown Law Journal,* 93 (2005): p. 427.

Tollefsen, D., "Organizations as True Believers," *Journal of Social Philosophy* 33 (2002): pp. 395–410.

Tuomela, R. *The Importance of Us: A Philosophical Study of Basic Social Notions* (Stanford, CA: Stanford University Press, 1995).

Velleman, D., "How to Share an Intention," *Philosophy and Phenomenological Research* 57 (1997): pp. 29–50.

NOTES

1. We thank Uriah Kriegel for his extremely helpful comments on the formulation of these distinctions.
2. Here we want to single out for special praise the work of Block 1978, 1995. His pioneering research in consciousness studies has deeply influenced the experimental studies presented below. In fact, the very term 'phenomenal consciousness' is borrowed from Block 1995, where he argues explicitly that people grasp the distinction between phenomenal and non-phenomenal states. Of course, Block's primary aim in those papers is somewhat different from our own, in that he is trying to use facts about people's intuitions as part of an inquiry into the true nature of consciousness.
3. For detailed methodological and statistical information, see Knobe and Prinz 2006.
4. We were surprised that subjects gave such high ratings for 'experiencing a sudden urge,' and we therefore ran a quick follow-up study to get a better handle on the phenomenon. Some subjects received the sentence 'Acme Corp. is experiencing a sudden urge to pursue internet advertising.' Others received a sentence that was exactly the same except that the word 'experiencing' was replaced with 'feeling.' The mean rating for subjects who received the version with 'experiencing' was 2.9; the mean for subjects who received the version with 'feeling' was 3.9. The overall mean was 3.5. These results suggest that the original ratings may have been artificially high as a result of sheer chance.
5. In a striking development, Michael Bruno, Bryce Huebner, and Hagop Sarkissian (unpublished data) have conducted a cross-cultural study demonstrating that this effect also arises among subjects in Hong Kong. (Even more interestingly, the study showed a significant effect such that the difference between ascriptions to groups and ascriptions to individuals is smaller for Hong Kong subjects than it is for American subjects.)
6. Here it is natural to wonder whether people would also be willing to ascribe phenomenal states to the corporation if it had been enchanted by a sorceress. We do not yet have any experimental data on this question, but Adam Arico and Shaun Nichols are designing a study to address the issue.
7. As the researchers rightly emphasize, phenomenal consciousness is specifically relevant to judgments of *moral patiency* (judgments about whether it would be wrong to do certain things to a given entity) rather than to judgments of *moral agency* (judgments about whether it would be wrong for the entity itself to do certain things). Judgments of moral agency appear to depend more on ascriptions of non-phenomenal states, such as beliefs, desires and intentions.

On Being an Octopus

Peter Godfrey-Smith

If octopuses did not exist, it would be necessary to invent them. I don't know if we could manage this, so it's as well that we don't have to. As we explore the relations between mind, body, evolution, and experience, nothing stretches our thinking the way an octopus does.

In a famous 1974 paper, the philosopher Thomas Nagel asked: What is it like to be a bat? He asked this in part to challenge materialism, the view that everything that goes on in our universe comprises physical processes and nothing more. A materialist view of the mind, Nagel said, cannot even begin to give an explanation of the subjective side of our mental lives, an account of what it feels like to have thoughts and experiences. Nagel chose bats as his example because they are not so simple that we doubt they have experiences at all, but they are, he said, 'a fundamentally *alien* form of life.'

From *Boston Review*, June 3, 2013. Reprinted by permission of the publisher.

Bats certainly live lives different from our own, but evolutionarily speaking they are our close cousins, fellow mammals with nervous systems built on a similar plan. If we want to think about something more truly alien, the octopus is ideal. Octopuses are distant from us in evolutionary terms, have a nervous system of very different design, and bodies with no bones and little fixed shape at all. What is it like to be an octopus? The question is intrinsically interesting and, beyond that, provides a good way to chip away at the problem Nagel raised for a materialist understanding of the mind.

• • •

How do we approach questions about 'what it's like' to be something or someone? One way of asking these questions makes them impossible to answer regardless of what minds might be made of. In this interpretation, to ask what it's like to be a bat or an octopus is to ask for a description, given from a third-person point of view, that encapsulates the animal's experience itself. But *having* an experience will always be different from having a *description* of it. This will be true if we are biochemical machines and true also if there is a soul-like extra ingredient in the world. A gap between a first-person and a third-person point of view arises either way. Descriptions are not completely powerless, though, in helping us get a grip on what the experience of another might be like. What a description can do, often very effectively, is prompt memories and guide the imagination—it can elicit memories of experiences that one has actually had and guide the construction of variations on these memories. Whenever one person describes an important experience to another, we rely on this sort of use of memory and imagination. It is more difficult if someone or something cannot talk, cannot offer a usable description in their own words. Then if we want to get a sense of what their experience might feel like, we must draw on information about their other forms of behavior and about how their senses and nervous systems work. If what is going on in them can be mapped onto what is going on in us when we have an experience that we know firsthand, we can say something about what an experience is like for them. Doing this does rely on the assumption that there is a systematic relation between how things feel and what goes on in the nervous system—just as listening to what someone says requires the assumption that real experiences lies behind her words.

The biologist Richard Dawkins offered a reply to Nagel's challenge about bats in his 1986 book *The Blind Watchmaker*. One of the main differences between bats and ourselves, emphasized by Nagel, is bats' use of sonar, sound pulses, for navigation. Nagel said that this was unlike any sense that we possess, and we cannot reach bat experience by imagining ourselves to have a more elaborate form of hearing. Dawkins replied that the use of sound in both human hearing and bat sonar is an incidental matter. Instead, using sonar as a bat would feel similar to the way seeing feels for us. We should not imagine sonar as upgraded hearing, but as modified seeing.

Dawkins based this claim on what sonar *does* for a bat. An animal that uses sonar constructs an internal model of the location of objects in space on the basis of stimuli from its environment. The information made available by sonar is not exactly the same as that made available by vision, and the resulting internal models will certainly differ, but, Dawkins argues, the feel of vision results from the way it enables you to make your way through the world, and that gives us some indication of what it feels like to navigate with sonar. This argument does not require that the same parts of the brain be used for each sense. Strikingly though, a 2011 brain-imaging study by Lore Thaler and her colleagues found that blind humans with some natural ability to echolocate using mouth-clicks were using parts of their brains normally dedicated to vision to process the clicks.

To work out what it might feel like to be another animal, we have to find some way to justify mappings between what goes on inside that animal and experiences that we can, through memory and imagination, partly conjure in ourselves. We do the same thing with humans who have different capacities and backgrounds from our own. Nagel accepted this claim about the human case but added that the more biologically distant the subject whose experiences we are inquiring into, 'the less success one can expect with this enterprise.' I agree that the more distant the subject, the more difficult the project becomes. But who's to say that success must remain elusive? On to the octopus.

• • •

An octopus has neurons, more or less like those of other animals, and many of them are organized into a brain. This brain evolved on an evolutionary path far removed from our own.

All animals have common ancestors if we go back far enough in time, and the pattern of relatedness between different animal species takes the rough form of a tree, the 'tree of life.' Our common ancestor with octopuses lies back near the beginning of the evolution of complex animals, perhaps 600 million years ago. That ancestor was a small, simple, marine animal—probably a flattened worm. Its many descendants include, on one branch, humans and the other animals with backbones (dolphins, bats, birds), and on another branch, a huge range of invertebrate animals, including the octopus.

The octopus's evolutionary path from the ancestral worm is unusual among the invertebrates because it led, as our path did, to a large nervous system. A common octopus has around 500 million neurons. That is many fewer than we have, but it is in the same range as a dog's brain, which has 600 million or so. The octopus, along with some of its cephalopod cousins, is an independent experiment in the evolution of a large nervous system, the only such experiment outside the vertebrates.

The result is an animal that is curious and a problem-solver. Some octopuses carry pairs of coconut half-shells around to reconstruct as spherical shelters. Octopuses can recognize (and take a disliking to) individual human keepers in aquariums. They learn the layout of their environment and hunt on long loops that take them reliably back to a den. Octopuses have eyes built on a 'camera' design like ours, with a lens focusing an image. They also have sensitive chemical sensors in their suckers—they taste the world as they touch it. When watching their eyes, it is natural to think that perhaps octopuses are a bit like us, just with more arms and no bones. Like other animals, they use their senses to track what is going on around them and to guide action. Would being an octopus be *so* different from being a bat, or any other animal with fine-tuned senses and a complex nervous system?

Underneath the skin, though, octopuses have an organization that takes them even further from us than they appear on the outside. Invertebrates generally have less centralized, more 'distributed' nervous systems than vertebrates such as us. Octopuses are mollusks (like oysters and clams), and their nervous systems are organized in part into ganglia, little knots of nerve cells, with links between the knots. Most mollusks do not have much of a central brain. Starting out from a molluskan layout of this kind, evolution increased the size of the octopus' nervous system enormously. The outcome of this process was uncovered in the mid-twentieth century, especially by John Z. Young and Martin Wells, working at the Naples Zoological Station in Italy. They found a number of surprises, some of which had consequences for all of biology. In 1936 Young described nerve cells in squid that are vastly larger than those of other animals, so big that electrodes could be inserted directly into them to measure their electrical activity. Work on those neurons became the basis for our understanding of how all nerve cells function.

Young, Wells, and their colleagues found that the octopus nervous system has three main parts. One is a central brain, which is a squashing-together of many expanded ganglia. There are also two optic lobes, large structures directly behind each eye. But most of the neurons—about two-thirds of them—are not in the head at all, but in the arms themselves. The connections between the central brain and this more peripheral nervous system also seemed to these researchers to be quite slim. An octopus's arms are packed with sensors responding to touch and chemistry, and the arms have enormous flexibility, able to bend in any direction at any point. So there is a lot going on in the arms, but the connections between arms and brain are apparently restricted to a narrow channel. From these anatomical facts and some experiments on behavior, the early researchers inferred that the arms have a good degree of independence from the central brain. They do their own sensing and their own responding. As Roger Hanlon and John Messenger summarized it in their 1996 book *Cephalopod Behavior*, the arms seemed 'curiously divorced' from the brain, at least with respect to the control of basic motions.

This disrupts a first round of guesses about what it might be like to be an octopus. Working from our understanding of their senses and the way they live their lives, it's natural to first imagine that the experience of an octopus is visually rich (though apparently in black-and-white) and augmented with elaborate chemical sensing. Everything touched by the arms is tasted. We can imagine something about what it might be like to live in this bright and tasty world. But then we realize that this line of thought might be fundamentally mistaken, as it assumes that an octopus is the same kind of psychological unit that a person is. It assumes that the locus of octopus experience is a psychological *self*, though perhaps a simple one, where the senses converge to generate a

feeling of how the world is. This picture might be wrong because an octopus is not organized as we are. Vision certainly feeds into the central brain, but each arm also contains shorter arcs between sensing and action.

This organization of the animal's control systems is the most difficult barrier to working out what octopus experience might be like. The best work I know of that bears on these issues is coming out of Benny Hochner's laboratory at Hebrew University in Jerusalem. In a 2011 study from Hochner's lab, Tamar Gutnick and her colleagues published a paper that looked at whether an octopus could guide a single arm along a complex maze-like path to a specific location to get food. The task was set up in such a way that the octopus could not merely let the arm's sensors follow a chemical gradient to the food, as the arm had to leave the water at one point to reach the target location. But the maze walls were transparent and the target location could be seen. To solve the problem the octopus had to guide its arm through the maze with vision. Although it took a while, all but one of the octopuses in the experiment learned to get an arm through to the food. The study also noted, though, that when octopuses are doing well with this task, the arm finding the food does what looks like its own local exploration at various stages, crawling and feeling around. There may be a mixture of two forms of control here: central control of the arm's general path and fine-tuning of the search by the arm itself. Another possibility is that, by means of attention of some kind, the octopus is exerting control over all the details of movements that might usually be more autonomous.

Suppose that the 'mixed-control' option, which Hochner tends to favor as an interpretation, is right. What would octopus experience be like? A range of partial analogies can be drawn with the human case. I visited Hochner's lab with another philosopher, Laura Franklin-Hall, who wondered: Would an octopus experience its arms more as parts of its environment than as straightforward parts of itself? The arms would not be experienced entirely as environment, because they can be centrally controlled to some extent—they are less 'divorced' from the brain than earlier researchers suspected. But once an arm has been sent in a certain direction, to some extent it is on its own. An analogy might be drawn with actions such as blinking or breathing. These are activities that normally happen involuntarily, but through attention you can assert control over them. The analogy is imperfect because although breathing is normally involuntary, when you do intervene to do it voluntarily, the control can be very fine-grained. In that case, attention is used to take over what is normally an automatic process. In the octopus, if the mixed-control interpretation is right, central guidance of the movements is never complete, and the peripheral system always has its say. Expressed too anthropomorphically, you would send an arm out deliberately and hope the local fine-tuning goes right.

Action by an octopus, then, would mix elements that are usually distinct in animals like us. When we act, the border between self and environment is usually fairly clear. When we move an arm, the arm can be controlled both in its general path and in the details. You can then watch your arm move, but what you are watching are the consequences of choices, or perhaps of habits that are the remnants of earlier choices. Various other things in the environment are not under your direct control at all, though they can be moved indirectly by manipulating them with your limbs. Uncontrolled movement by an object around you is usually a sign that it is not part of you at all (with partial exceptions for knee-jerk reflexes and the like). If you were an octopus, these distinctions would be blurred. Your arms would move in a way that is a mix of the centrally and peripherally controlled. To some extent you would guide them, and to some extent you would just watch them go.

One might wonder whether the guided action seen in the Gutnick experiment is a normal behavior for an octopus or instead something entirely artificial. When searching for food, an octopus often puts all its arms around or under a rock and seems to just let them roam. If the experimental behavior is unusual, that would not make the work uninteresting. The fact that the octopus *can* solve the problem—can pull itself together in this way—would still be significant. However, the behaviors in the experiment may not be that unnatural in any case. I once saw an octopus searching for food at a boulder underneath which a large shark was resting. The octopus held its body well back and stretched one arm under the boulder, very long and straight, and seemed to watch its path closely.

There is also a more directly skeptical response to these ideas about octopus experience and the self-environment relationship. I've assumed up to this point that it makes sense to think that an octopus could have a feeling of agency, a feeling that tracks the difference between what it is controlling and what is merely

happening. But perhaps this is so sophisticated a form of experience that it is beyond any non-human animal? Although animals do act, perhaps they cannot feel *that* they are doing so; the contrast between actions and other events would not be apparent to them.

For at least some animals, this is probably not true: they may indeed have awareness of a distinction between events they cause and those they do not. This has been the topic of a number of interesting recent experiments. In these studies chimps or monkeys first learn to play simplified video games, moving virtual objects by using a joystick, trackball, or another controller, trying to get certain objects to meet or collide and, in some cases, trying to get other objects to avoid collision. They then perform tasks of these kinds while another 'distracter' object moves on the screen in similar ways to the object they are controlling. The objects then freeze and the chimps' second task—the one the experiment is set up to study—is to indicate which object was moving under their control.

In a 2011 study by Takaaki Kaneko and Masaki Tomonaga, chimps did well on a task of this kind. They guided an object toward a moving target and then picked it out from a distracter whose motions were those of an object that had been guided by a chimp on an earlier trial. It's reasonable to wonder if chimps are a special case here, but Justin Couchman has done a related series of experiments on rhesus monkeys. Couchman's experiment had the monkeys doing a harder task than the chimps were faced with, and one of his four monkeys clearly mastered it.

Every report of this kind I have read raises interesting further puzzles. Kaneko and Tomonaga ran an additional experiment in which the chimps were controlling *neither* icon as it moved. Instead they were watching a recording of an entire earlier trial. Moving the trackball had no effect on any object. As expected, the chimps did better at choosing the right object when they were actually controlling it. But surprisingly, they did not do too badly—performed better than chance—even when neither object was under their control. How is this possible? What does it even mean to get the 'right' answer when neither object is being controlled? The 'right' object was the one that had been controlled in the earlier trial when it was recorded. The object being controlled at that time will tend to follow the target more closely than the object that had been a distracter in that trial. This difference in apparent goal-directedness might lead the chimps to be more inclined to choose that object. Given this, it is important that the chimps made better choices when they had real control of one of the objects.

Nothing like these experiments has been tried in octopuses, as far as I know, and a chimp or a monkey is a very different animal from an octopus. But the experiments certainly tell against the idea that awareness of agency is beyond all non-human animals.

Some philosophers working within a broadly materialist framework are opposed to asking questions about 'what it's like' to have a particular kind of mind. They regard this way of setting up the issues as misguided. I think these questions are good ones, as long as they are asked in a way that does not doom them to unanswerability from the start. The divide between first-person and third-person points of view is real regardless of what minds are made of. Knowing how an animal's body and brain are put together does not put you into a state that is similar to what is going on inside that animal, so in that sense no description can tell you 'what it's like to be' that animal. Getting a sense of what it feels like to be another animal—bat, octopus, or next-door neighbor—must involve the use of memory and imagination to produce what we think might be faint analogues of that other animal's experiences. This project can be guided by knowledge of how the animal is put together and how it lives its life. When the animal is as different from us as an octopus, the task is certainly difficult, but it is one worth undertaking. Doing so is part of the attempt to strike a balance between treating our minds as too private and mysterious to make scientific sense of at all, and treating them as less private and mysterious than they really are.

The Self

The problems of the self are multifarious. First, what is the self? Second, what is self-consciousness, the awareness of self? Third, what is it for a self or a person to persist over time. These three interconnected issues are often called the problem of the self, the problem of self-consciousness, and the problem of personal identity.

Chapter 66 by Galen Strawson introduces these issues, focusing on the sense of self: the sense which he says all normal human beings have of being a self. Strawson sets out our unreflective concept of the self as a single mental thing, a subject of experience which is distinct from all other things and which has its own personality. He argues that personality is inessential to being a self, as is the sense of oneself as a single continuous being across time. He suggests that one's sense of self can be episodic rather than narrative, without flow from one moment to the next.

One traditional view of the self is the dualist view: the self is a primitive nonphysical entity distinct from the body and the brain, which in principle could exist without the body and the brain. That view is covered in the first five chapters of the book: Gyekye on the Akan view of persons, Avicenna on self-consciousness, and Descartes and Elisabeth of Bohemia on mind-body dualism. Those chapters are also highly relevant here.

Another traditional view of the self is that the self does not exist. The Buddha famously said that there is no self, and this view has been at the core of Buddhist philosophy since then. As a result, the Buddhist tradition has reflected longer and harder on issues about the self than any other. Mark Siderits (chapter 67) discusses the Buddhist conception of self and personal identity, clarifying the doctrine of no-self and drawing out Buddhist arguments for the doctrine.

If there is a self, what is it? A common view is that one's self is one's brain. Andy Clark's piece "I am John's Brain" (chapter 68) brings out the many ways that the self and the brain seem to differ, from the brain's point of view. Patricia Churchland (chapter 69) argues for a close connection between the brain and the self, suggesting that the self is in effect a construction of the brain. On Churchland's view, understanding neurobiology is the key to understanding the self.

Derek Parfit (chapter 70) discusses the question of personal identity: what makes a person the same person over time. If we were immaterial Cartesian egos, then being the same self would be a matter of being the same ego. But if we are not, then the issue of personal identity must be a matter of physical and psychological continuity over time. He considers cases involving teletransportation, where some or all of one's body and one's psychology is replaced. Is the teletransported being you? Parfit argues that sometimes

these questions have no answers, and that we need to revise our views about personal identity as a consequence.

Greg Egan's story "Learning to Be Me" (chapter 71) centers on a puzzling scenario involving personal identity and the self. Everyone has a "jewel" in their head, a sort of backup system that learns to function exactly like the brain. It seems that there are two selves present in every skull. Any given person can wonder whether they are the brain or the jewel. Or perhaps, while the brain and the jewel are synchronized, there is only one self? If the jewel and the brain become unsynchronized, does one person turn into two? When people switch to a mode where only the jewel is present and not the brain, does the original person then die? Any theory of the self can be tested on the rich puzzles that this story provides.

Susan James (chapter 72) discusses personal identity from a feminist perspective. She argues that much work on personal identity has understated the role of the body in personal identity. She focuses especially on cases of brain transplantation, where many theorists have argued that psychological continuity matters more than bodily continuity. James argues that in separating these sorts of continuity and privileging the psychological over the bodily, theorists privilege the symbolically masculine over the symbolically feminine. On a feminist view, the body is more important to identity than these psychological views allow.

In recent years, "identity" has often been understood as referring to *social identity:* the socially significant categories such as gender, race, and nationality with which one identifies. The concept of social identity is not the same as the concept of personal identity in the preceding chapters, but it is perhaps equally important to understanding the self and even more important in our lives. Kwame Anthony Appiah (section 73) offers an analysis of many dimensions and complexities of social identity and the roles that it plays in our lives. He also argues that behind many social identities there lie important philosophical assumptions that may turn out to be false, including an essentialist view of the self on which our identities are part of our true nature or essence.

FURTHER READING

Appiah 2018, Churchland 2013, Parfit 1984, and Strawson 2009 defend their views of the self and identity at length. Perry 1978 is a classic dialogue on personal identity. Schechtman 1996 defends an account of the self-grounded in the autobiographical narratives we produce. Olson 1999 defends the view that we are animals. Dainton 2008 defends a view of personal identity grounded in consciousness. Ganeri 2012 offers a perspective on the self-grounded in Indian philosophy. Alcoff 2005 addresses social identity with special attention to gender and race.

Alcoff, L. M. *Visible Identities: Race, Gender, and the Self* (New York: Oxford University Press, 2005).

Appiah, A. *The Lies that Bind: Rethinking Identity* (New York: W. W. Norton, 2018).

Churchland, P. S. *Touching a Nerve: The Self as Brain* (New York: W. W. Norton, 2013).

Dainton, B. *The Phenomenal Self* (Oxford: Oxford University Press, 2008).

Ganeri, J. *The Self: Naturalism, Consciousness, and the First-Person Stance* (Oxford: Oxford University Press, 2012).

Olson, E. *The Human Animal: Personal Identity Without Psychology* (Oxford: Oxford University Press, 1999).

Parfit, D. *Reasons and Persons* (New York: Oxford University Press, 1984).

Perry, J. *A Dialogue on Personal Identity* (Indianapolis: Hackett, 1978).

Schechtman, M. *The Constitution of Selves* (Ithaca, NY: Cornell University Press, 1996).

Strawson, G. *Selves: An Essay in Revisionary Metaphysics* (New York: Oxford University Press, 2009).

The Sense of the Self

Galen Strawson

By the 'sense of the self' I mean the sense that people have of themselves as being, specifically, a mental presence, a mental someone, a conscious subject that is distinct from all its particular experiences, thoughts, hopes, wishes, feelings, and so on. This core sense of self comes to every normal human being, in some form, in childhood (it is not a 'Western' peculiarity). It is perhaps most often vivid when we are alone and thinking, but it can be equally vivid in a room full of shouting people. It connects with a feeling that most people have had at some time, that their body is just a vehicle or vessel for the mental thing they really are. (Neither physical activity nor pain need diminish our sense of the independence of the self from the body; they're just as likely to increase it.) I'm not claiming that the sense of the self automatically incorporates belief in an immaterial soul, or in life after bodily death. It doesn't. Philosophical materialists have as strong a sense of self as anyone else, although they believe, as I do, that we are wholly physical beings, and that human consciousness evolved by purely physical processes.

Our natural, unreflective conception of the self seems to have at least six main elements. First, the self is thought of as a thing, in some sense. Second, it is thought of as specifically mental, in some sense. Third, it is thought of as single. Fourth, it is thought of as something that has a certain character or personality. Fifth, it is thought of as something that is in some manner distinct from all other things. Sixth, it is thought of as something that is a subject of experience, a conscious feeler, thinker, chooser. In considering each element in turn, I use the expression 'the self' freely, as a loose name for all the undeniably real phenomena that lead us to think and talk in terms of the self. This doesn't rule out the possibility that the best thing to say, in the end, is that there is no such thing.

The first claim, that the self is thought of as a thing, is, in a way, the least clear. The general idea is this: it isn't thought of as a state or property of something else, or as an event, or as a mere process or series of events. To this extent, there is nothing else for it to seem to be, other than a thing—not a thing in the way that a stone or a chair is, some sort of ethereal concrete object, but a thing of some kind. It is thought of, in particular, as something that has the causal character of a thing; something that can undergo and do things. Bishop Berkeley's characterisation of the self as a 'thinking active principle'[1] seems as good as any. In this old use, a principle manages to sound like a thing without sounding like a table or a chair.

The second claim, that the self is something mental, is also unclear, but the central idea is this: when the self is thought of as a thing, its claim to thinghood is taken to be sufficiently grounded in its mental nature alone; it may also have a non-mental nature, as materialists suppose, but its being a thing is not thought to depend on its counting as a thing considered in its non-mental nature. The self is the mental self. It's true that people naturally think of themselves as possessing both mental and non-mental properties, but this doesn't affect the standard conception of the mental self.

Although experience of the mental self needn't involve any belief in an immaterial soul, it does incorporate elements that make that belief come rather naturally. The mental self can easily seem to exist self-sufficiently in a sphere of being quite other than that described by physics. Things are not as they seem, according to materialists; but they certainly seem as they seem, and this helps to explain how natural it is to think of the self as a specifically mental thing.

The third claim is that the self is thought of as single. But in what way? Not as a single assemblage or collectivity, as a pile of marbles is single, but rather, as a marble is single when compared with a pile of marbles. Furthermore, it is standardly thought to be single both when it is considered synchronically, i.e. as existing at a given time, and when it is considered diachronically, i.e. as a thing that persists through time. I will take 'synchronic' to apply to any consideration of the self during what is experienced as a unitary or 'hiatus-free' period of consciousness. 'Diachronic' will then apply to any period that includes a break or hiatus. Truly

From *London Review of Books*, vol. 18, no. 8, 18 April 1996; slightly revised. Reprinted by permission of the publisher.

unbroken periods of experience are, I suggest, almost invariably brief: a few seconds at most, a fraction of a second at the least.

I've said that we normally have a sense of the singleness of the self. Some, however, may claim to experience it as fragmentary or multiple, and most of us have had experience that gives us—we feel—some understanding of what they mean. In fact there are reasons for thinking that the experience of multiplicity can only really affect the sense of the self diachronically considered, not synchronically. But this may be doubted: we may be subject to rapidly changing and overlapping moods, and conflicts of desire. Our thought processes can become extraordinarily rapid and tumultuous, disparate contents tumbling over one another. This may be claimed to involve experience of the self as synchronically multiple.

To experience conflicting desires is not to experience the self as multiple, however. On the contrary, it's one of the most vivid forms of experiencing oneself as single; as a single thing that feels conflict precisely because it is single. What about the mind racing chaotically? When this happens many people experience themselves as helpless spectators of the pandemonium, which is more likely to increase than to diminish their sense of being single. Furthermore, any supposed experience of the self as synchronically multiple will have to be an episode of explicitly self-conscious thought; but there is a crucial sense in which such experience is incompatible with genuine self-conscious thought. 'The subjective 'I' can never be . . . divided, and it is this 'I' that we presuppose in all thinking,' as Kant remarks.[2] Don't believe anyone who tells you otherwise, especially if they come from Paris.

The fourth claim is simply that the self is thought of as having character or personality, in exactly the same way as an embodied human being. This is hardly surprising, for we take it that our personality is a matter of how we are mentally speaking; so if we think that our existence involves a mental self, we're bound to think that that self has a personality.

The fifth claim is that the self is thought of as ontologically distinct. From what? The question has various answers. To begin with, the self is thought of as ontologically distinct from any of its conscious mental goings-on—thoughts, experiences, and so on. It *has* thoughts and experiences, but is not the same as them, or constituted out of them. A stronger version of this view holds that the self is distinct not only from any conscious mental goings-on, but also from any non-conscious mental features like beliefs, preferences, stored memories, character traits. Hume famously challenged the first of these views, proposing that a persisting self, if it exists at all, may be a series of mental goings-on. Ordinary thought rejects this 'bundle' theory, however, as Hume later did himself, and endorses the second, stronger version.

A third and still stronger version rejects materialism, claiming that the self is ontologically distinct from anything physical. But this dualist, or idealist, idea is not an integral part of the sense of the self.

The sixth claim, that the self is a subject of experience, I take to be obvious. What is a subject of experience? The ordinary notion seems pretty clear: independently of any metaphysical commitments, each of us has a very good idea of what a subject of experience is just in being one and being self-conscious. The self is clearly not the only thing that is thought of as a subject of experience: it's just as natural (or more natural) for us to say that a human being considered as a whole is a subject of experience—as are millions of non-human animals. Nevertheless, we have a tendency to think that in the human case it is above all the mental self that is the subject of experience.

What about the view that the self is capable of action, in thinking and choosing, for example? Many would wish to add a seventh, separate, claim—that the self is thought of as an agent, as 'the source of effort and attention, and the place from which . . . emanate the fiats of the will' in William James' words.[3] This seems very plausible.

Suppose these seven elements capture the conceptual core of the ordinary human sense of the self. Are they all essential to anything that could count as a genuine sense of the self? I will challenge the claims about personality and diachronic singleness.

The principal point to be made against the personality claim is simple. We already have a natural way of conceiving of the self according to which it does not have a personality, but is, strictly speaking, a mere 'locus' of awareness. Most people have at some time, and however temporarily, experienced themselves not just as neutral and unengaged, but as stripped of particularity of character: as a mere point of view. This may be the result of exhaustion or solitude, or it may just be how some people feel when they wake up. It may be the temporary result of abstract thought or a hot bath. It is also a

common feature of severe depression, in which we experience 'depersonalisation'—an accurate term. Depersonalisation is pathological, but it is experientially real, and one can imagine getting stuck in this condition. (Some people do.) Equally, one can imagine a race of alien beings for whom it is the normal condition, but who still have a clear sense of the mental self as the locus of consciousness.

A very strong form of what may be lost in depersonalisation is recorded by Gerard Manley Hopkins, who talks of considering

> my self-being, my consciousness and feeling of myself, that taste of myself, of I and me above and in all things, which is more distinctive than the taste of ale or alum, more distinctive than the smell of walnut leaf or camphor, and is incommunicable by any means to another man . . . Nothing else in nature comes near this unspeakable stress of pitch, distinctiveness, and selving, this self-being of my own.[4]

I find this bewildering. I find it quite hard to believe that Hopkins is telling the truth and have yet to meet someone whose experience resembles his. For most people, their personality is something unnoticed, and in effect undetectable, in the present moment. It's what they look through, or where they look from; not something they look at.

It is harder to dislodge the idea that a genuine sense of the self must incorporate a conception of it as something that has relatively long-term diachronic singleness or continuity. Yet that sense may be vivid and complete, at any given time, even if it has to do only with the present brief, hiatus-free stretch of consciousness. It may be said that although this is a formal possibility, it is remote from reality and from our interests: that life without any significant sense of the long-term continuity of the self is conceivable for aliens, but hardly for ourselves. Strictly speaking, all I need for my argument is the formal possibility; but it seems to me that life without any such sense of long-term continuity lies well within human experience. We can be fully aware that we have long-term bodily continuity without having any such sense of the self as something persisting. The idea may have very little—or no—emotional importance for us. It may contribute little or nothing to the overall character of our experience. Human beings differ deeply in a number of ways that may affect their experience of continuity.

In considering these differences, I sometimes write 'in the first person' like William James, 'leaving my description to be accepted by those to whose introspection it may commend itself as true, and confessing my inability to meet the demands of others, if others there be.'[5] Clearly James doesn't really believe that there are others unlike himself, any more than Hume does when he pretends to allow that some people may perceive a simple and continuous mental self when they introspect, although he is 'certain there is no such principle in me.'[6] My position is different. When it comes to the differences I am about to discuss, I believe that there are others quite unlike myself.

First, there are differences of memory. Some people have an excellent 'personal' memory (as philosophers call memory of one's own past) and an unusual capacity for vivid recollection. And their personal memory may not be just reliable and retentive: it may also be highly active and regularly intrude into their present thoughts. Others have a very poor personal memory, which may also be quiescent, and almost never intrude. These differences of memory are matched by equal differences in the force with which people imagine, anticipate or form intentions about the future.

These differences interact with others. Some people live in narrative mode, and wrongly assume that everyone else does the same: they experience their lives in terms of something that has shape and story, a narrative trajectory. Some of them keep diaries with posterity in mind and imagine future biographies. Some are self-narrators in a stronger sense: they regularly rehearse and revise their interpretations of their lives. Some are great planners and knit up their lives with long-term projects. Others have no early ambition, no later sense of vocation, no interest in climbing a career ladder, no tendency to see their life as constituting a story or development. Some merely go from one thing to another, living life in a picaresque or episodic fashion. Some people make few plans and are little concerned with the future. Some live intensely in the present, some are simply aimless. This can be a basic fact of character or the outcome of spiritual discipline; it can be a response to economic destitution—a devastating lack of opportunities—or vast wealth. There are lotus-eaters, drifters, lilies of the field, mystics, and people who work hard in the present moment. There are many possibilities. Some people are creative although they lack ambition or long-term aims and go from one small thing to the next, or produce large works without planning to, by accident or accretion. Some people are very consistent in character, whether or not

they know it, a form of steadiness that may underwrite experience of the self's continuity. Others are consistent in their inconsistency and feel themselves to be continually puzzling and piecemeal. Some go through life as if stunned.

I'm somewhere down towards the episodic end of this spectrum. I have no sense of my life as a narrative with form and very little interest in my own past. My personal memory is very poor and rarely impinges on my present consciousness. I make plans for the future and to that extent think of myself perfectly adequately as something with long-term continuity. But I experience this way of thinking of myself as remote and theoretical, given the most central or fundamental way in which I think of myself, which is as a mental self or someone. Using ME to express the way in which I think of myself, I can accurately express my experience by saying that I do not think of ME as being something in the future.

It is January as I write this. The thought that I have to give a Wolfson College lecture in March causes me some anxiety, and this has familiar physiological manifestations. I feel the anxiety naturally and directly as pertaining to me even though I have no sense that it will be ME that will be giving the lecture. Indeed it seems plain false to say that it will be ME. And this is how it feels. It's not just something I happen to believe for theoretical reasons. So why do I feel anxiety? Doubtless because my susceptibility to it is innate and 'hard-wired' connected with the instinct for self-preservation: my concern for my future, which is within the normal human range, is biologically grounded and autonomous in such a way that it persists as something immediately felt even though it is not supported by any emotionally backed sense on the part of ME now that ME will be there in the future.

My experience of the self is just one kind among others; no doubt some people have it in a more extreme form. It matters here only insofar as it supports the claim that a sense of the self need not necessarily involve experience of it as something with long-term continuity. This experience may be common, but it is not universal, it fades over time in some and is withered, in others, by reflection.

Some think that conscious experience flows, that this is simply given prior to any theoretical suppositions. According to James, 'Consciousness . . . does not appear to itself chopped up in bits. Such words as 'train' or 'chain' do not describe it fitly as it presents itself in the first instance. It is nothing jointed; it flows. A 'river' or a 'stream' are the metaphors by which it is most naturally described . . . Let us call it the stream of consciousness, or of subjective life.'[7] This seemed a good move in 1890, given the dominant psychological atomism that inspired the metaphors of trains and chains, collections, bundles, and heaps. But perhaps we have now been misled in the opposite direction, into thinking consciousness more fluent than it is. This is important: for if consciousness does feel stream-like then this may be part—although only a part—of the explanation of why many people have a sense of continuity. In fact I think the metaphor of the stream is inapt, even though streams contain pools and falls—not to mention weeds and stones. Thought has very little natural continuity or experiential flow—if mine is anything to go by. It keeps slipping from mere consciousness into self-consciousness and out again. It is always shooting off, shorting out, spurting, and stalling. James likened it to 'a bird's life . . . an alternation of flights and perchings'[8] (the idea is beautifully developed), but even this image retains strong continuity insofar as a bird traces a spatio-temporally continuous path, and it fails to take adequate account of the fact that trains of thought are constantly interrupted by detours, fissures, by blows and white noise.

This is especially so when we are just sitting thinking. Things are different if our attention is engaged by some ordered and continuous process in the external world, like a fast and exciting game, or music. In this case thought or experience may be felt to take on the ordered continuity of the phenomenon which occupies it. But it may still cut out and restart, or flash with extraneous matter from time to time, and it is arguable that the case of solitary speculative thought merely reveals in a relatively dramatic way something that is true to a greater or lesser extent of all thought. Some will think that James Joyce's use of full stops in *Ulysses* makes his depiction of consciousness more accurate in the case of Leopold Bloom and Stephen Dedalus (many full stops) than in the case of Molly Bloom (none; some will find that her great flood of words resembles speech more than thought). There may be some difference between the sexes—Virginia Woolf claimed that Dorothy Richardson had 'invented the psychological sentence of the feminine gender'[9]—but I doubt that they are significant.

Radical disjunctions in the process of consciousness don't occur only at the level of content. Switches of subject-matter can be absolute

and still be seamless in that they involve no sensible temporal gap or interruption of consciousness. It seems to me that the experience of seamlessness is relatively rare. When I'm alone and thinking and consider my thinking I find that the fundamental experience is one of repeated returns into consciousness that suggest an immediately prior state of complete, if momentary, non-consciousness. The (invariably brief) periods of true continuity are usually radically disjunct even when it is the same thought (or nearly the same thought) that I return to after an episode of absence. It is as if consciousness is continually restarting. It keeps banging out of nothingness; it is a series of comings-to.

Some hiatuses involve complete switches of focus and subject-matter. Others occur between thoughts that are connected in subject-matter, or when we are attending to something in such a way that we hardly notice the hiatus because the content of experience is more or less the same afterwards as before. In this case, the hiatus may be a mere caesura, an entirely accidental feature of the mechanism of consciousness. But it is likely that it is also functional in some way, part of a basic process of regirding attention: a new 'binding' of the mental manifold, a new synthesis in the Kantian sense. The hiatus is often fast: it's not hard to overlook the absolute fugues and interstitial vacancies of consciousness—just as we overlook the blinks of our eyes. But they are easily noticeable when attended to, available to memory in our current state of consciousness.

Perhaps this is a rash generalisation from my own case, or an unwitting confession of schizophrenia. I think, though, that careful mental self-examination may reveal the same to everyone, if in different degrees. That said, the facts are hard to access, and belief that the process of consciousness is a matter of flow may itself contribute to an experience of flow.

But perhaps the experience of disjunction is an artefact of introspection: perhaps the facts get distorted by the attempt to observe them. Perhaps unexamined consciousness has true flow. The reply to this objection is, first, that even if the appearance of disjunction were partly—or even largely—an artefact, this would be a striking fact about how consciousness appears to itself, important when considering the underpinnings of a sense of continuity. A second reason the objection seems wrong is that awareness of disjunction can surface spontaneously: we can become aware that this is what has been happening, rather than seeing it happen only when we look. In a sense, the issue is undecidable, for in order to decide it we would need to be able to observe something while it was unobserved.

Insofar as it finds support in the moment-to-moment nature of consciousness, then, the sense of continuity does not derive from a phenomenon of steady flow, but from other sources—such as the constancies and coherences of content that often link up experiences through time, and, by courtesy of short-term memory, across the radical jumps and breaks of flow. I work in a room for an hour; I look up at the rain on the window and turn back to the page; I hold the same pen throughout. Examined in detail, the processes of my thought may be scatty. And yet I am experientially in touch with a great pool of constancies and steady processes of change in my environment, which includes my body. These constancies and steadinesses in the contents of consciousness may seem like fundamental characteristics of its operation, although they are not. And this in turn may support the sense of the self as something truly continuous throughout the waking day and smooth the path to the idea of it as an entity that may be continuous also during sleep, and so from week to week to month to year.

Although I have no sense of seamless flow in the process of consciousness, it doesn't immediately follow that I have no sense of the continuing mental self. When we first try to think about our sense of the self—and to think about it, rather than simply have it, is already difficult—our first reaction may well be that it does present the self as a single thing continuing throughout the waking day: something that has all the interrupted thoughts and experiences but is not itself interrupted. This reaction is backed up by our awareness of our continuity as embodied human beings with robustly persisting sets of basic beliefs, preferences, mental abilities, and so on. In my case, however, the reaction is weak and soon undermined. As I think further about my mental life, I'm met by the sense that there is no 'I' that goes on through the waking day (and beyond). I feel I have continuity only as an embodied human being. If I consider myself as a mental subject of experience, my sense is that I am continually new.

I don't mean new or different in respect of personality and outlook. I have a perfectly adequate grasp of the similarities that characterise me from day to day. But when I consider the fundamental experience of myself as a mental self, my feeling is that I am continually new. In his autobiography, John Updike writes: 'I have

the persistent sensation, in my life . . . that I am just beginning.'[10] This seems exactly right. The experience of the 'I' as in some sense new each time is (I suggest) fundamental and universally available, although it is occluded for many by familiar and contrary habits of thought, and may emerge clearly only on reflection. I feel I'm a nomad in time, although the metaphor is intussusceptive, because it is the 'I' itself that has the transience of abandoned camping-grounds.

Research by Pöppel shows that the 'conscious now' is about three seconds long: this is the most we can hold together at any one time, experientially speaking.[11] 'In this sense,' writes Miroslav Holub, 'our ego lasts three seconds.'[12] His claim is tangential to mine. I don't think that the brevity of the 'conscious now' necessarily contributes to the sense of hiatus or newness. Our experience could resemble a narrow beam of light sweeping smoothly and continuously along. The length of the 'conscious now' may set an upper limit on hiatus-free periods of thought, but it doesn't follow that there will always be conscious experience of hiatus within any four-second period (there may be none for days). Nor, crucially, am I claiming that the self will never appear to last longer than the 'conscious now,' when reflected on. My use of the word 'long-term' is vague but intentionally so: the self can certainly be felt to persist throughout a period of time that includes a break or hiatus, and its temporal extent may appear very different in different contexts of thought (fear of death raises interesting questions).

Some may doubt my claims about how I experience consciousness; those who do not may think I'm part of a small minority. Experience like mine may be thought to be the unnatural result of doing philosophy, or drugs. But even if the experience of disjunction was specifically the result of philosophical reflection, it wouldn't follow that it resulted from philosophy distorting the data. Philosophy may simply make us examine the already existing nature of our experience more closely. Even if the experience were unnatural or uncommon in daily life, it would not follow that it gives a less accurate picture of how things are; for many natural experiences represent things inaccurately. More important, the experience may be natural in the sense that any ordinary human being who considers the matter will find that he or she comes to have it.

Human beings, then, can have a vivid sense of the self without having any sense of it as something that has either personality or long-term continuity. Does this improve the prospects for the claim that a sense of the self could be an accurate representation of something that actually exists—even if materialism is true? I think it does, although the full argument would require a careful statement of what it is to be a genuine, realistic materialist (a realistic materialist is fully realist about consciousness), and, further inquiry into the notion of a thing, and a challenge to the problematic distinction between things including processes. Perhaps the best account of the existence of the self is one that may be given by certain Buddhists. It allows that the self exists, at any given moment, while retaining all the essential Buddhist criticisms of the idea of the self. It gives no reassurance to those who believe in the soul, but it doesn't leave us with nothing. It stops short of the view defended by many analytic philosophers, according to which the self is a myth insofar as it is thought to be different from the human being considered as a whole. It leaves us with what we have, at any given time—a self that is both materialistically respectable, and distinctively mental, and as real as a stone.

NOTES

1. *Three Dialogues between Hylas and Philonous*, 1713, ed. J. Dancy (Oxford 1998), 116.
2. I. Kant, *Critique of Pure Reason*, 1781–87, trans. N. Kemp Smith (London Macmillan 1933), A354.
3. W. James, *The Principles of Psychology*, 1890 (Harvard 1981), 285
4. G. M. Hopkins, "Commentary on the Spiritual Exercises of St Ignatius Loyola," 1880, in *Sermons and Devotional Writings* (Oxford 1959), 123.
5. W. James *The Principles of Psychology*, 1890 (Harvard 1981), 286.
6. D. Hume, *Treatise Of Human Nature*, 1739–40, ed. D. F. Norton and M. Norton (Oxford 2000), §1.4.6.3.
7. W. James *The Principles of Psychology*, 1890 (Harvard 1981), 233.
8. Ibid., 236.
9. V. Woolf, "A Romance of the Heart," 1923, in *The Essays of Virginia Woolf*, vol. 3 (London: Hogarth Press, 1987), 367.
10. J. Updike, *Self-Consciousness* (London: Deutsch, 1989), 239.
11. See e.g. E. Pöppel, "Time Perception," in *Handbook of Sensory Physiology*, vol. 8 (New York: Springer 1978), 713–729; and E. Ruhnau, "Time-Gestalt and the Observer," in *Conscious Experience*, ed. T. Metzinger (Thorverton: Imprint Academic 1995), 165–84.
12. M. Holub, *The Dimension of the Present Moment* (London: Faber 1990), 5.

Non-Self: Empty Persons (Excerpt)

Mark Siderits

The Buddha holds that we experience the suffering of *saṃsāra* because of our ignorance of the three characteristics: impermanence, suffering, and non-self. Of these three, it is the characteristic of non-self that plays the central role in his diagnosis. According to early Buddhism, there is no self, and persons are not ultimately real. This may be put somewhat cryptically as: we are empty persons, persons who are empty of selves. In this chapter we will investigate this claim. We will look at some of the arguments found in early Buddhist texts for the claim that there is no self. And we shall try to determine what it means to say that persons are not ultimately real. But before we can do either of these things we need to determine what it would mean to say that there is a self. The word 'self' gets used in several different ways, only one of which is relevant to the philosophical question the Buddha is trying to answer. We can avoid much confusion about what Buddhists mean by their doctrine of non-self if we begin by getting clear concerning what they mean when they speak of a self.

3.1

By 'the self,' what Buddhists mean is the essence of a person—the one part whose continued existence is required for that person to continue to exist. This is the definition of 'self' that we will use. But what does it mean? It might be helpful to think of the view that there is a self as one possible answer to the question what it is that the word 'I' refers to. I am a person. And persons are made up of a variety of constituents: parts making up the body, such as limbs and organs, and parts making up the mind, such as feelings and desires. Now persons are things that continue to exist for some time—at least a lifetime, if not longer. But not all the parts of a person must continue to exist in order for that person to continue to exist. I could survive the loss of a finger or toe. And I might lose my desire for coffee without ceasing to exist. So apparently not all the parts of a person are necessary to the continued existence of a person. To say there is a self is to say that there is some one part that is necessary. This one part would then be what the word 'I' really named. The other parts would more properly be called 'mine'; only that one essential part would count as the true 'me.' The alternative to this would be to say that 'I' refers to all the parts collectively. Let us call this alternative the view that 'I' is the name of the person, where by 'person' we mean the whole that consists of all the parts that make up my body and mind over the duration of my existence. So either 'I' is the name of some one essential part of the person or else it refers to the person as a whole. (Of course this applies to the other words we use to refer to persons as well, such as names.)

To say there is a self is to say that there is some one part of the person that accounts for the identity of that person over time. If there were a self, then the person whose self it was would continue to exist as long as that self continued to exist. The self would then be the basis of a person's identity over time. It would be what explained why this present person, me, is the same person as some earlier person. But we need to be careful with the expression 'same person.' For the English word 'same' is ambiguous. When we say '*x* and *y* are the same,' there are two things we might mean. We could mean that *x* and *y* are qualitatively identical, or we could mean that *x* and *y* are numerically identical. To say that *x* and *y* are qualitatively identical is to say that they share the same qualities, that they resemble one another or are alike. To say that *x* and *y* are numerically identical is to say that they are one and the same thing, that '*x*' and '*y*' are really just two names for one entity. So there can be cases of qualitative identity but numerical distinctness, as with two T-shirts that come out of the factory looking exactly alike. And there can also be cases of numerical identity but qualitative distinctness, as with a leaf that in summer is green and smooth but in autumn is red and crinkled. We said above that according to the self-theorist, a self is what explains why some person existing

From *Buddhism as Philosophy: An Introduction* (Ashgate, 2007). Reprinted by permission of the publisher.

now is the same person as someone who existed earlier. The key thing to keep in mind is that here 'same' is meant in the sense of numerical identity.[1]

Like many other things, persons can undergo very significant qualitative changes and yet continue to exist. I can continue to exist as one and the same person, me, even though the properties I now have are quite different from those I used to have. Thanks to the ambiguity of the English word 'same,' we can put this as, 'He is the same person but not the same.' When we say this we are not contradicting ourselves. The first 'same' ('the same person') is used in the sense of numerical identity. The second 'same' is used in the sense of qualitative identity; 'not the same' means qualitatively distinct. It is one person, me, who once had the property of liking coffee, but now has the very different property of disliking coffee. A person can undergo qualitative change while retaining numerical identity. Since the self is supposed to be what explains numerical identity over time of persons, perhaps a self could undergo qualitative change. What it could not undergo is numerical change, that is, going out of existence and being replaced by another self.

If there is a self, it is 'what makes me me,' 'the true me,' that which 'gives me my identity.' These ways of describing what a self is are all open to a common misinterpretation. People often speak of 'discovering their self,' of 'finding their true identity.' What they often mean by this is figuring out which characteristics seem most important or valuable. So someone might say that they have come to realize their identity isn't tied up with physical appearance but with less superficial things like artistic talent or communication skills. Discoveries like this are probably important to personal growth. But they have nothing to do with what the Buddhists mean by a self. We can see this from the fact that even if there is no self, we can still ask which of a person's characteristics are most important to that person's happiness. To speak of a self is to speak of some one part of the person, the part that must always exist as long as the person exists. To speak of an 'identity' that can be 'found' is to speak of characteristics or properties, of what a person is like. There might very well be no single part of the person that must continue to exist in order for that person to continue to exist. (This is exactly what the Buddha is going to argue for.) But it might still be true that some characteristics of a person play a more important role in their life than others. Otherwise it wouldn't make sense to say that a person has 'lost their identity.' Perhaps my life would be less meaningful if I were to lose those traits that now have great importance to me. But it would still be my life. I could survive that qualitative change. I might be a very different kind of person. But I would still be me.

There is another misinterpretation that arises in connection with the idea that the self is what gives me my 'identity.' It is common to think that someone's identity is what sets that person apart from all others. Add to this the idea that one's identity consists in what one is like, one's characteristics or properties. The result is the notion that a self would be what makes one different from everyone else. Now the word 'different' is ambiguous in the same way that 'same' is: there is numerical difference or distinctness, and there is qualitative difference. If it's numerical distinctness that is meant, then it's true that the self would be what makes one different from others. If we have selves, then my self and yours must be two distinct things, not one. But it's not true if what's meant by 'different' is qualitative difference. It is not true that if we had selves, each would have to be unique in the sense of being unlike every other. Two selves could be perfectly alike, like two peas in a pod, and still serve to make one person numerically distinct from another.

The difficulty with the idea that the self must be qualitatively unique is that it once again confuses the notion of the self with the notion of what one is like, one's properties or characteristics. And properties may be shared between two things, whereas numerical identity may not. The leaf on this branch of this tree today might be exactly like the leaf that was here last year—same color, same shape, same pattern of veins, etc. But they are numerically distinct leaves all the same. Perhaps no two persons are ever exactly qualitatively alike. Even twins who share DNA patterns have physical differences, such as different fingerprints. Still there is no contradiction involved in supposing that there might be two persons who are exactly qualitatively alike. Imagine for instance that each of us has lived countless lives in the past. Given the innumerably many beings there may be in the universe, it does not seem unlikely that someone somewhere might once have lived a life just like the one I am now living. Yet that would have been someone else, not me. So if what makes me the person I am is my self, then my self is not what makes me qualitatively unlike other people.

Suppose, moreover, that each person is qualitatively unlike every other. This could be true even if there were no selves. Indeed it could be true if there were selves that were all qualitatively identical. This is actually something that many non-Buddhist Indian philosophers hold. On their view, the self is something that is simple or impartite (lacking parts). The self is just the subject of experiences, the part of us that is aware of the different experiences we have. Your self and mine would then be just like those two peas in a pod. It's common to suppose that what makes different people qualitatively different is that they have different experiences. But on this view of the self, the different experiences that people have would not make their selves qualitatively different. Since the self is simple, it cannot be changed by the experiences it is aware of. It is other parts of the person that are changed by those experiences. The experience of eating changes the shape of my body. The experience of smelling coffee changes a desire in my mind. My self is unaffected by these changes, it is simply aware of them. Someone holding this view of the self who also thought that persons are qualitatively unique could say that their uniqueness is explained by facts about those parts of the person that are not the self. Someone who denied the existence of a self could explain the qualitative uniqueness of persons in the same way.

3.2

In order to show that the self does not exist, we need to know what we are looking for, and where to look. We now know that a self would be that part of the person that 'I' is consistently used to refer to. So we can tell what to look for by seeing how we actually use words like 'I.' For instance, we say things like 'I was born in New York, now live in the Midwest, and will move to Arizona when I retire.' So if 'I' refers to the self, the self would have to be some one numerically identical thing that continues to exist throughout the past, present and future history of the person. There are more clues to be found in the ways we use this word, but this tells us enough for present purposes. Where should we look? Since the self is supposed to be a part of the person, we obviously need to look among the parts that make up persons. It would be helpful if we had a list of the basic categories of person-parts. This is just what the Buddha provides with his doctrine of the five *skandhas.* (The word *skandha* is here used in its sense of 'bundle.') These are:

- *Rūpa:* anything corporeal or physical;[2]
- Feeling: sensations of pleasure, pain and indifference;
- Perception: those mental events whereby one grasps the sensible characteristics of a perceptible object; e.g., the seeing of a patch of blue color, the hearing of the sound of thunder;
- Volition: the mental forces responsible for bodily and mental activity, for example, hunger, attentiveness, and
- Consciousness: the awareness of physical and mental states.

A word of caution is necessary concerning these categories. Their names are here being used as technical terms, with precise definitions. Do not confuse these with the ordinary meanings of these words. For instance, the second *skandha*, feeling, refers only to the three kinds of hedonic sensation: pleasure, pain and indifference (neither pleasure nor pain). It does not include most of the things that are often called 'feelings,' such as the emotions of anger and jealousy. Those emotions go under the very different *skandha* of volition. Likewise by 'consciousness' is here meant just the awareness itself, and not what it is that one is aware of. So when I am conscious of a pain sensation, there are two *skandhas* involved: the pain, which goes under feeling *skandha*, and the consciousness that is aware of it, which goes under consciousness *skandha.* Again, we sometimes use the word 'perception' to refer to our beliefs about and attitude toward something. So someone might say, 'My perception of the new government is that it is weak and will soon fall.' This is not the sort of thing that would go under perception *skandha.* This is a complex mental state, whereas an instance of perception *skandha* is a simple mental event. A perception in this technical sense is just the occurrence of a sensory content to the mind: the simple thought of a patch of blue or the smell of lemon.

The five *skandhas* are sometimes referred to collectively as *nāma-rūpa* (sometimes translated as 'name and form'). Here *nāma* refers to the four *skandhas* other than *rūpa.* The literal meaning of *nāma* is 'name,' but here it means 'that which can only be named.' The idea is that while *rūpa* can be perceived by the external senses, the members of the four other categories cannot be seen or touched. Because they are not publicly observable, we cannot explain

what they are by pointing; we can only communicate about them through the names we have learned to use for these private states. What this tells us is that the doctrine of the five *skandhas* expresses a kind of mind-body dualism. The Buddha is claiming that in addition to those parts of the person that we can see and touch—the parts of the body—there are other constituents that are not themselves physical. Some philosophers today hold the view called 'physicalism,' according to which all that exists is physical. On this view there is no more to a person than the physical constituents, their body and brain. What we think of as mental events, such as thoughts and emotions, are really just complex brain events. When the Buddha says that in addition to *rūpa skandha* there are the four *nāma skandhas*, he is in effect denying that physicalism is true. On his account, mental events are separate non-physical kinds of things. We will be looking at this claim more carefully later on.

The Buddha uses the doctrine of the five *skandhas* as a tool in his search for a self. He goes through each *skandha* in turn and tries to show that nothing included in that category could count as a self. But this raises a new question: would this really show that there is no self? Isn't it possible that the self exists elsewhere than among the five *skandhas*? In order for the Buddha's strategy to work, he will have to show that the doctrine of the five *skandhas* gives an exhaustive analysis of the parts of the person. We will call this the 'exhaustiveness claim.'

The exhaustiveness claim is this: every constituent of persons is included in one or more of the five *skandhas*.

In the following passage, the later commentator Buddhaghosa argues in support of this claim.

> The basis for the figment of a self or of anything related to a self, is afforded only by these, namely *rūpa* and the rest. For it has been said as follows:
>
> When there is *rūpa*, O monks, then through attachment to *rūpa*, through engrossment in *rūpa*, the persuasion arises, 'This is mine; this am I; this is myself.'
>
> When there is feeling . . . when there is perception . . . when there are volitions . . . when there is consciousness, O monks, then through attachment to consciousness, through engrossment in consciousness, the persuasion arises, 'This is mine; this am I; this is my self.'
>
> Accordingly he laid down only five *skandhas*, because it is only these that can afford a basis for the figment of a self or of anything related to a self.
>
> As to other groups which he lays down, such as the five of conduct and the rest, these are included, for they are comprised in volition *skandha*. Accordingly he laid down only five *skandhas*, because these include all other classifications. After this manner, therefore, is the conclusion reached that there are no less and no more. [VM xiv.218]

This at least makes clear that Buddhists recognize the need to support the exhaustiveness claim. But it is not clear how good an argument this is. The idea seems to be that these are the only things we are aware of when we are aware of persons and so come to believe that persons have selves. Is this true? And if it were true, would it show that the exhaustiveness claim is true? We will return to this question.

3.3

Let us now look at how the Buddha formulates his arguments for non-self. In the following passage the Buddha is addressing his five former companion *śramanas*. . . . It contains two distinct arguments. The first is what we will call the argument from impermanence, since it is based on the claim that all five *skandhas* are impermanent or transitory. But there is also a second argument here.

> Then The Blessed One addressed the band of five *sramanas:*
>
> '*Rūpa*, O monks, is not a self. For if now, O monks, this *rūpa* were a self, then this *rūpa* would not tend towards destruction, and it would be possible to say of *rūpa*, 'Let my *rūpa* be this way; let not my *rūpa* be that way!' But inasmuch, O monks, as *rūpa* is not a self, therefore does *rūpa* tend towards destruction, and it is not possible to say of *rūpa*, 'Let my *rūpa* be this way; let not my *rūpaa* be that way!'
>
> 'Feeling . . . perception . . . volitions . . . consciousness, is not a self. For if now, O monks, this consciousness were a self, then would not this consciousness tend towards destruction, and it would be possible to say of consciousness, 'Let my consciousness be this way; let not my consciousness be that way!' But inasmuch, O monks, as consciousness is not a self, therefore does consciousness tend towards destruction, and it is not possible to say of consciousness, 'Let my consciousness be this way; let not my consciousness be that way!'

'What do you think, O monks? Is *rūpa* permanent, or transitory?'

'It is transitory, Reverend Sir.'

'And that which is transitory—is it painful, or is it pleasant?'

'It is painful, Reverend Sir.'

'And that which is transitory, painful, and liable to change—is it possible to say of it: 'This is mine; this am I; this is my self'?'

'Certainly not, Reverend Sir.'

'Is feeling . . . perception . . . volition . . . consciousness, permanent, or transitory?'

'It is transitory, Reverend Sir.'

'And that which is transitory—is it painful, or is it pleasant?'

'It is painful, Reverend Sir.'

'And that which is transitory, painful, and liable to change—is it possible to say of it: 'This is mine; this am I; this is my self'?'

'Certainly not, Reverend Sir.'

'Accordingly, O monks, as respects all *rūpa* whatsoever, past, future, or present, be it subjective or existing outside, gross or subtle, mean or exalted, far or near, the correct view in the light of the highest knowledge is as follows: 'This is not mine; this am I not; this is not my self.'

'As respects all feeling whatsoever . . . as respects all perception whatsoever . . . as respects all volitions whatsoever. . . as respects all consciousness whatsoever, past, future, or present, be it subjective or existing outside, gross or subtle, mean or exalted, far or near, the correct view in the light of the highest knowledge is as follows: 'This is not mine; this am I not; this is not my self.'

'Perceiving this, O monks, the learned and noble disciple conceives an aversion for *rūpa*, conceives an aversion for feeling, conceives an aversion for perception, conceives an aversion for volitions, conceives an aversion for consciousness. And in conceiving this aversion he becomes divested of passion, and by the absence of passion he becomes free, and when he is free he becomes aware that he is free; and he knows that rebirth is exhausted, that he has lived the holy life, that he has done what it behooved him to do, and that he is no more for this world.'

Thus spoke The Blessed One, and the delighted band of five *śramanas* applauded the speech of The Blessed One. Now while this exposition was being delivered, the minds of the five *śramanas* became free from attachment and delivered from the depravities. [S III.66–68]

Here the Buddha cites two different sorts of reasons why the *skandhas* are not the self: they are impermanent ('subject to destruction,' 'transitory'), and they are not under one's control ('painful,' 'it is not possible to say of *x*, 'Let my *x* be this way . . ."). To separate out the argument from impermanence from the second argument, let's ignore the claims about the five *skandhas* not being under one's control (we'll discuss this in section 4), and focus on the claims about their being subject to destruction and transitory. If we add the exhaustiveness, claim as an implicit premise,[3] the argument is then:

1 *Rūpa* is impermanent.
2 Sensation is impermanent.
3 Perception is impermanent.
4 Volition is impermanent.
5 Consciousness is impermanent.
6 If there were a self it would be permanent.

IP (There is no more to the person than the five *skandhas*.)
C Therefore there is no self.

This argument is valid or logically good. That is, if the premises are all true, then the conclusion will also be true. So our job now will be to determine if the premises really are all true. But before we can do that, there is one major point that needs clarifying: just what do 'permanent' and 'impermanent' mean here? Once again the doctrine of karma and rebirth becomes relevant. For those like the Buddha and his audience who accepted the doctrine, 'permanent' would mean eternal, and 'impermanent' would mean anything less than eternal. This is because if we believe it is the self that undergoes rebirth, and we also believe that liberation from rebirth is possible, then we will hold as well that the self is something that continues to exist over many lives, and can even exist independently of any form of corporeal life. This is probably what the Buddha had in mind with premise 6. And in that case, all that would be needed to show that something is not a self is to establish that it does not last forever—even if it did last a longtime. So if, for instance, the *rūpa* that is my body does not last forever, then it is not my self. And of course my body does go out of existence when I die, so this would be sufficient to show that it is not my self.

What about those of us who do not accept the doctrine of karma and rebirth? To believe in rebirth is to believe that the person exists both before and after this life. If we do not believe in rebirth, then we may believe that the person exists only a single lifetime. In that case, a self would not have to exist any longer than a lifetime in order to serve as the basis of a person's numerical identity over time. So all that

'permanent' in premise 6 could mean is 'existing at least a whole lifetime.' It could not mean 'eternal.' Likewise, to show that a *skandha* is impermanent in the relevant sense, we would have to show that it does not exist for the entire duration of a person's life. Does this mean that the argument won't work without the assumption of karma and rebirth? After all, isn't it true that our bodies last for our entire lives?

Not necessarily. First, we need to remember that the self is supposed to be the essential part of the person, and the body is a whole made of parts. Which of these parts—the organs that make up the body—is the essential one? There doesn't seem to be any single organ that I could not live without. Granted I could not survive without a heart. But as heart replacement surgery shows, I don't need *this* heart in order to continue to exist. If my heart were my self, then when I got a replacement heart I would cease to exist and someone else would then be living in my body. That replacement heart came from someone else, so it would be that person's self. But surely if I chose to have heart replacement surgery I would not be committing suicide! What about the brain? Not only can I not live without a brain; there is no such thing as brain replacement surgery, so I cannot live without this brain. But here the problem seems to be entirely practical, not an 'in-principle' difficulty. If we knew how to reprogram an entire brain, then we might be able to replace a diseased brain with a healthy one while preserving all of a person's psychology. This would be like copying the contents of the failing hard drive of your computer, replacing the hard drive, then reinstalling everything onto the new hard drive.

This brain-replacement scenario might seem too science-fictional to support premise 1. But there's a second reason someone might give for denying that the body is permanent in the relevant way. This is that all the parts of the body are constantly being replaced—at the level of the molecules that make up our cells. We've all heard it said that none of the atoms that made up our body seven years ago is among those making up our body now. Life processes such as metabolism and meiosis involve the constant, piecemeal replacement of the parts that make up a life-form. After these processes have gone on long enough, all the matter making up a given organ is new: the atoms now making up that organ are numerically distinct from the atoms that made it up earlier. Given this, it could be said that the body and brain I have now are not numerically identical with the body and brain I had seven years ago. ***Rūpa*** would then be impermanent in the relevant sense.

We have been discussing how to interpret premise 6, the premise that a self would have to be permanent, and how premise 1, which says that ***rūpa*** is impermanent, might be true in light of our interpretation of 6. Our general practice in examining arguments will be to first look at what reason there might be to think that the premises are true, and then to evaluate the argument overall. How might someone defend the remaining premises, 2–5? These are not affected by the question of karma and rebirth in the way that premise 1 is. For regardless of whether we interpret 'permanent' to mean eternal, or just to mean lasting a single lifetime, the four *nāma skandhas* will all count as impermanent. This is the point the Buddha makes in the following passage:

> It would be better, O monks, if the uninstructed worldling regarded the body which is composed of the four elements as a self, rather than the mind. And why do I say so? Because it is evident, O monks, that this body which is composed of the four elements lasts one year, lasts two years, lasts three years, lasts four years, lasts five years, lasts ten years, lasts twenty years, lasts thirty years, lasts forty years, lasts fifty years, lasts a hundred years, and even more. But that, O monks, which is called mind, intellect, consciousness, keeps up an incessant round by day and by night of perishing as one thing and springing up as another.
>
> Here the learned and noble disciple, O monks, attentively considers dependent origination: 'this exists when that exists, this originates from the origination of that; this does not exist when that does not exist, this ceases from the cessation of that.' O monks, a pleasant feeling originates in dependence on contact with pleasant objects; but when that contact with pleasant objects ceases, the feeling sprung from that contact, the pleasant feeling that originated in dependence on contact with pleasant objects ceases and comes to an end. O monks, an unpleasant feeling . . . an indifferent feeling originates in dependence on contact with indifferent objects; but when that contact with indifferent objects ceases, the feeling sprung from that contact, the indifferent feeling that originated in dependence on contact with indifferent objects ceases and comes to an end.
>
> Just as, O monks, heat comes into existence and flame into being from the friction and concussion of two sticks of wood, but on the separation and parting of these two sticks of wood the heat sprung from those two sticks of wood ceases and comes to an end; in exactly

> the same way, O monks, a pleasant feeling originates in dependence on contact with pleasant objects; but when that contact with pleasant objects ceases, the feeling sprung from that contact, the pleasant feeling that originated in dependence on contact with pleasant objects, ceases and comes to an end. An unpleasant feeling . . . an indifferent feeling originates in dependence on contact with indifferent objects; but when that contact with indifferent objects ceases, the feeling sprung from that contact, the indifferent feeling that originated in dependence on contact with indifferent objects ceases and comes to an end. [S 11.96f]

Of course the Buddha knows that reflective people are more likely to consider the mind the self than the body. In the Western philosophical tradition this is just what Descartes did. He concluded that the true 'I' is not the body but the mind—a substance that thinks (that is, is conscious), endures at least a lifetime, and is immaterial in nature. Many Indian philosophers reached somewhat similar conclusions. The Buddha's point is that the conclusion that the mind endures at least a lifetime rests on an illusion. For what we call the mind is really a continuous series of distinct events, each lasting just a moment, but each immediately followed by others. There is no such thing as the mind that has these different events, there are just the events themselves. But because they succeed one another in unbroken succession, the illusion is created of an enduring thing in which they are all taking place.

The eighteenth-century British philosopher David Hume said something similar in response to Descartes. Descartes claimed to be aware of the mind as something that is aware, that cognizes, perceives, wills, believes, doubts—that is the subject of all one's mental activities. Hume responded that when he looked within, all he ever found were particular mental contents, each of them fleeting, and never an enduring substance that has them. He concluded that it is just the relations among those mental events that make us invent the fiction of the self as an enduring subject of experience. The Buddha claims something similar. And like Hume, he uses the relation of causation to support his claim.

In the last chapter we saw how the doctrine of dependent origination is used to explain the origin of suffering. In the passage we are looking at, that doctrine gets put to a different use. Dependent origination is the relation between an effect and its causes and conditions. Where this relation holds, the effect will arise when the causes and conditions obtain, and the effect will not occur when the causes and conditions do not. The Buddha asserts that all the *nāma skandhas* are dependently originated. He uses the example of feeling, but this example generalizes to the other kinds of mental events as well. Consider the feeling of pleasure I derive from eating my favorite kind of ice cream. This feeling originates in dependence on contact between my sense of taste (located in the taste buds on my tongue) and the ice cream. Before that contact there was no feeling of pleasure, and when the contact ceases so does the feeling. I may have a feeling of pleasure in the next moment, but that occurs in dependence on a new event of sense-object contact—say, when I take my next bite of ice cream. So that feeling is numerically distinct from the first, for it has a different cause. One feeling has gone out of existence and been replaced by another. Now the senses are by nature restless, always making contact with new objects. This means that there will be an unbroken stream of feelings and other mental events. It is easy to mistake this stream for a single enduring thing. But the Buddha claims that if we attend to the individual events making up this stream, then seeing how they are dependently originated will help us overcome the illusion of a persisting subject of experience.

The appeal to dependent origination is meant to show two things: that there is no such thing as the mind over and above the mental events making up the mental stream; and that each of those events is very short-lived. Suppose we agreed with the Buddha on the first point. How successful is this appeal with regard to the second point? It is relatively easy to agree that feelings of pleasure and pain are transitory. We don't really need to use dependent origination to prove this. And since they are transitory, they could not be the self. Likewise for perceptions. But what about volitions? Granted my desire for some new soft drink may last only as long as the effects of the commercial I just saw. But we also seem to have volitions that endure, such as my desire for coffee. To this it could be replied that this is an acquired volition, one that I did not always have and might very well get rid of. So the opponent must look for volitions that seem to endure a whole lifetime. They might suggest what are sometimes called 'instinctual desires,' such as the desire to escape

life-threatening situations. Might this not be a volition that is permanent in the relevant sense? The Buddha will reply that what we are then describing is not one enduring volition, but rather a pattern of recurring volitions, each lasting only a brief while before ceasing. This is shown by the fact that I am only aware of a desire to escape danger when I perceive a threatening situation. The desire thus originates in dependence on a specific sense-object contact event, and ceases to exist when that event ceases. The opponent will then want to know what explains the pattern of recurring volitions. What the opponent suspects is that this pattern can only be explained by supposing that there is one enduring volition, a permanent desire to escape life-threatening situations, that is always present in me. My perception of a life-threatening situation brings the volition out into the part of my mind that is illuminated by consciousness, but it persists even when I am not aware of it.

Since we have no evidence that the Buddha was ever presented with this line of objection, we don't know how he would have responded. But later Buddhist philosophers do show us how it might be answered. What we have here is a certain phenomenon—a pattern of recurring desires over the course of a person's lifetime—and two competing theories as to how to explain the phenomenon. Call the opponent's theory the 'in-the-closet' theory, since it claims that some desires continue to exist hidden away in a dark corner of the mind when not observed. It explains the phenomenon by claiming that it is a single continuously existing volition that manifests itself at different times as the desire to duck a falling safe, the desire to dodge a runaway car, etc. The Buddhist dependent origination theory, by contrast, claims that these are many numerically distinct desires. It explains the pattern by appealing to the ways in which the parts of a person's body are arranged. Consider the thermostat that controls the heat in a house. It is because of the way in which the parts of the thermostat are put together that whenever the temperature goes below a certain threshold, the thermostat signals the furnace to go on. It is not as if the signal for the furnace to go on waits in the thermostat's closet until the room gets too cold. By the same token, the Buddhist would say, it is because of the way that the human body is organized that a danger stimulus causes a danger-escaping volition.[4] Now this seems like a plausible explanation. It makes sense to suppose that, for instance, it is because of the way in which certain neurons in the brain are arranged that we have this desire to escape whenever we sense danger. But the in-the-closet theory also seems plausible to many people, so which should we choose?

There is a principle that governs cases like this. It is known in the West as Ockham's Razor, but Indian philosophers call it the Principle of Lightness, for it dictates that we choose the 'lighter' of two competing theories. The Principle of Lightness may be stated as follows: given two competing theories, each of which is equally good at explaining and predicting the relevant phenomena, choose the lighter theory, that is, the theory that posits the least number of unobservable entities.

To posit an unobservable entity is to say that something exists even though we never directly observe that thing. Now you might think that positing an unobservable entity is always a bad idea. Why believe something exists when no one can see or feel it? But modern physics tells us that there are subatomic particles like electrons and protons, and no one has ever seen or felt such things. Does that make modern physics an irrational theory? No. What the Principle of Lightness tells us is that we should only posit unobservable entities when we have to, when there is no other way to explain what we observe. We accept the theory that says there are subatomic particles because no other theory does as good a job of explaining the phenomena. In the case of the phenomenon of recurring desires, though, things are different. We said that the in-the-closet theory and the Buddhist dependent origination theory give equally good explanations of this phenomenon. But the in-the-closet theory posits an unobservable entity that the dependent origination theory does not. The former theory says that volitions continue to exist in our minds even when we are not aware of them. The latter theory speaks instead of patterns of neurons in the brain—something that can be observed. This makes the latter theory lighter, and so it is the theory that we ought to choose.

The Principle of Lightness would help the Buddhist answer the objection about seemingly permanent volitions. It can also be used in defense of premise 5, the premise that says consciousness is impermanent. In the following passage the Buddha claims that consciousness also originates in dependence on sense-object contact:

> Just as, O monks, fire is named from that in dependence on which it burns. The fire which

> burns in dependence on logs of wood is called a log-fire. The fire which burns in dependence on chips is called a chip-fire. The fire which burns in dependence on grass is called a grass-fire. The fire which burns in dependence on cow-dung is called a cow-dung fire. The fire which burns in dependence on husks is called a husk-fire. The fire which burns in dependence on rubbish is called a rubbish-fire. In exactly the same way, O monks, consciousness is named from that in dependence on which it comes into being. The consciousness which comes into being in respect of color-and-shape in dependence on the eye is called eye-consciousness. The consciousness which comes into being in respect of sounds in dependence on the ear is called ear-consciousness. The consciousness which comes into being in respect of odors in dependence on the nose is called nose-consciousness. The consciousness which comes into being in respect of tastes in dependence on the tongue is called tongue-consciousness. The consciousness which comes into being in respect of things tangible in dependence on the body is called body-consciousness. The consciousness which comes into being in respect of *dharmas* in dependence on the mind is called mind-consciousness. [M 1.259–60)]

To this someone might object that we experience consciousness as some one thing that endures. That when I first see and then take a bite of ice cream, it is one and the same consciousness that is first aware of the color of the ice cream and is then aware of the taste of the ice cream. The Buddhist would respond by pointing out that there are periods in a person's life when there seems to be no consciousness at all occurring. If the opponent were to claim that consciousness continues to exist even then—only in the closet—the Buddhist could reply that their theory of dependent origination gives a lighter explanation of the apparent continuity of consciousness.[5]

But the Principle of Lightness would also help the Buddhist defend their claim that the mind is an invented fiction. As both the Buddha and Hume point out, we are never actually aware of the mind as something standing behind such mental events as feeling, perceiving and willing. We are just aware of the feelings, perceptions, and volitions themselves. So the mind is unobservable. And it is the causal relations among these mental events that the Buddha says explain all the facts about our mental lives. So the mind becomes an unnecessary, unobservable posit.[6]

Why, though, should we accept the Principle of Lightness? The idea behind this principle is that what makes some statement true has to be objective: the truth of a statement is not determined by such subjective factors as our interests, or limitations in our cognitive capacities, but rather just by facts that are independent of our interests and limitations. The thought is that when it comes to finding out what the facts are, we should let the world outside our mind dictate what it is that we believe. To think that factors in my mind could determine what the facts are would be to indulge in magical thinking. By the same token, we could say that positing unobservable entities is inherently suspect. Why believe that something exists when no one could possibly observe it? Because saying so makes it easier for us to explain what we do observe? This is letting what seems to us like a good explanation determine what we say the mind-independent facts are. This is letting our cognitive limitations determine what statements we believe are true. Magical thinking. The Principle of Lightness says we should resort to positing unobservable entities only when the world tells us we have no alternative.

We are now done with our review of the explicit premises in the argument from impermanence. There still remains the one implicit premise, the exhaustiveness claim. If we accept this, then it seems we must say the argument from impermanence succeeds in establishing that there is no self. There is one important objection to the exhaustiveness claim. Many find this claim unacceptable because it leaves unexplained the sense we have that there is an 'I' that has a body and various mental states. If the exhaustiveness claim were true, then while there would be a body and various mental states such as feelings and desires, these would not be the body and mental states of anyone or anything. They would be ownerless states without a subject. And this strikes many as absurd. Is this a valid objection to the exhaustiveness claim, and so to the argument from impermanence? We will defer this question.

Oh? I want to know more

NOTES

1. The ambiguity of 'same' is often resolved by context. When we say '*x* and *y* are the same P,' what is meant is numerical identity. When we say '*x* and *y* are the same,' what is often meant is qualitative identity. So I might say that this is *the same leaf* as the one I showed you yesterday, meaning that they are

one and the same leaf. Or I might say that this leaf is *the same* as the one that was on this branch last year, meaning that the two leaves are qualitatively identical. Other languages lack this ambiguity. In German, for instance, one says *das selbe* for numerical identity, and *das gleiche* for qualitative identity.

2. The literal meaning of *rūpa* is 'form' or 'shape,' and you will sometimes see the word rendered as 'form' in translations of Buddhist texts. But as the name of the first *skandha, rūpa* actually means 'that which has form or shape,' that is, anything material or physical. This is one case where it's best to slick with the Sanskrit original rather than try to come up with an acceptable English translation.
3. An implicit premise is an unstated premise that must be supplied for an argument to work, and that the author of the argument did not state because they thought it would be redundant—typically because it seemed to the author to be common knowledge that the author and the audience shared. We will follow the practice of putting implicit premises in square brackets.
4. No Buddhist text actually says this. This represents an extrapolation from what members of the Sautrāntika school of Abhidharma say about continuity of karmic seeds during meditational states in which there is no consciousness. Their approach to that problem is dictated by their overall aversion to talk of dispositions or powers as real things.
5. The Buddha's argument in the passage we just looked at is slightly different. It depends on the claim that there are six distinct kinds of consciousness, corresponding to the six senses and their respective objects. These twelve items (vision and the visible, hearing and the audible, etc.) are collectively referred to as the *āyatanas*.
6. Remember, though, that early Buddhism is dualist. One can deny the existence of the mind and still be a dualist. The most familiar form of dualism is substance dualism, the view that there are two kinds of substance, physical substance and mental substance. Descartes was a substance dualist. Buddhists deny the existence of the mind. But they affirm the existence of mental events, such as feeling and perception, as things that are distinct from the physical (*rūpa*). While early Buddhism denies substance dualism, it affirms what could be called event dualism.

I am John's Brain[1]

Andy Clark[2]

I am John's[3] brain. In the flesh, I am just a rather undistinguished looking grey/white mass of cells. My surface is heavily convoluted, and I am possessed of a fairly differentiated internal structure. John and I are on rather close and intimate terms; indeed, sometimes it is hard to tell us apart. But at times, John takes this intimacy a little too far. When that happens, he gets very confused about my role and functioning. He imagines that I organize and process information in ways which echo his own perspective on the world. In short, he thinks that his thoughts are, in a rather direct sense, my thoughts. There is some truth to this of course. But things are really rather more complicated than John suspects, as I shall try to show.

In the first place, John is congenitally blind to the bulk of my daily activities. At best, he catches occasional glimpses and distorted shadows of my real work. Generally speaking, these fleeting glimpses portray only the products of my vast subterranean activity, rather than the processes which give rise to them. Such products include the play of mental images or the steps in a logical train of thought or flow of ideas.

John's access to these products is, moreover, itself a pretty rough and ready affair. What filters into his conscious awareness is somewhat akin to what gets on to the screen display of a personal computer. In both cases, what is displayed is just a specially tailored summary of the results of certain episodes of internal activity: results for which the user has some particular use. Evolution, after all, would not waste time and money (search and energy) to display to John a faithful record of inner goings-on unless they could help John to hunt, survive, and reproduce. John, as a result, is appraised of the bare minimum of knowledge about my inner activities. All he needs to know is the overall significance of the upshot of a select few of these activities: that part of me is in a state which is associated with the presence of a

From *Journal of Consciousness Studies* 2:144–8, 1995. Reprinted by permission of the publisher.

dangerous predator and that flight is therefore indicated, and other things of that sort. What John gets from me is thus rather like what a driver gets from an electronic dashboard display: information pertaining to those few inner and outer parameters to which the gross activity of the agent can make a useful difference.

John, however, begs to differ. He thinks this is a crazy parallel since in his case there is no further agent to be informed by any 'dashboard display.' There is no 'driver' apart from me, his brain. But despite this undoubted fact, I insist that there is a dashboard display of sorts. The display consists of those select products of my activities which are able to play a role in those projects and decisions which the world at large ascribes to John-the-person (as opposed to those, like the maintenance of blood flow, ascribed not to John's decisions, but to John-the-biological-organism). The dashboard display thus consists of those products of my activity which are able to figure in what other humans would identify as John's plans, his choices and projects. Thus if one of my many sub-systems is appraised of some item of information, that item may or may not become available to support John's conscious planning and deliberate action. Information which is made available for such purposes can, of course, figure in John's on-going reflections on his own life and goals, while the rest, though often vital for John's continued success, remains invisible to John-the-agent. The fact that John has only limited access to my operations means, of course, that John can sometimes be unaware of the true causes of his own actions. In such cases, John is driven to create complex stories or narratives which try to make sense of his self-observed behaviours. This is a hard task, since the roots of much of that behaviour lie, I am proud to report, in those other activities of mine to which John has no conscious access. As a result, his stories are often wildly imaginative (that is to say, false) attempts to make sense of his own activities on the restricted basis of the 'dashboard display' types of information.

And it gets worse. For John's reports, even of the favoured 'dashboard display' products of my activity, are themselves filtered through the distorting lens of John's biased and limited vocabulary for reporting these facts to others. Thus John thinks (falsely) that introspection reveals the presence of entities he calls 'beliefs,' others he calls 'desires,' still others he calls 'hopes' and so on and so on. John is even inclined (in more philosophical moments) to picture these putative inner entities as sharing the basic structure of the very sentences he would use to report such facts to others. He thinks he finds in himself the belief that Rome is pretty, and the hope that St. Louis is pretty. And just as these sentences share a word 'pretty,' so John believes the internal states which 'carry' the thoughts must share a component too. I do not know why John thinks this, although at times he has such a loose idea of a 'component' that what he says cannot help but be true. I assure you, however, that on any non-trivial reading, what he says is false. John should beware of confusing the structure of the language he uses to report his beliefs with the structure of my own encodings. I like to store information in ways which make my unseen labours easier and which come naturally given my evolutionary history—a proud and long one for most of which the recent fad of language-use had not even been invented. My modes of information storage and retrieval, I can safely say, bear no deep resemblance whatever to these new-fangled linguistic vehicles with which John is so misleadingly familiar.

A further complex of misapprehensions centres on the question of the provenance of thoughts. John thinks of me as the point source of the intellectual products he identifies as his thoughts. But, to put it crudely, I do not have John's thoughts. John has John's thoughts, and I am just one item in the array of physical events and processes which enable that thinking to occur. John is an agent whose nature is fixed by a complex interplay between a mass of internal goings-on (including my activity) and a particular kind of physical embodiment and a certain embedding in the world. The combination of embodiment and embedding provides for persistent informational and physical couplings between John and his world; couplings which leave much of John's 'knowledge' out in the world and available for retrieval, transformation and use as and when required.

Take a simple example. A few days ago, John sat at his desk and worked rather hard for a sustained period of time. Eventually he got up and left his office, satisfied with his day's work. 'My brain,' he reflected (for he prides himself on his physicalism), 'has done very well. It has come up with some neat ideas.' John's image of the events of the day depicted me as the point source of those ideas; ideas which he thinks he captured on paper as a mere convenience and a hedge against forgetting. I am, of course, grateful that John gives me so much credit. He attributes the finished intellectual products directly

to me. But in this case, at least, the credit should be extended a little further. My role in the origination of these intellectual products is certainly a vital one: destroy me and the intellectual productivity will surely cease! But my role is more delicately constituted then John's simple image suggests. Those ideas of which he is so proud did not spring fully formed out of my activity. If truth be told, I acted rather as a mediating factor in some rather complex feedback loops encompassing John and selected chunks of his local environment. Bluntly, I spent the day in a variety of close and complex interactions with a number of external props. Without these, the finished intellectual products would never have taken shape. My role, as best I can recall, was to support John's re-reading of a bunch of old materials and notes, and to react to those materials by producing a few fragmentary ideas and criticisms. These small responses were stored as further marks on paper and in margins. Later on, I played a role in the re-organization of these marks on clean sheets of paper, adding new on-line reactions to the fragmentary ideas. The cycle of reading, responding and external re-organization was repeated again and again. At the end of the day, the 'good ideas' (with which John was so quick to credit me) emerged as the fruits of these repeated little interactions between me and the various external media. Credit thus belongs not so much to me as to the spatially and temporally extended process in which I played a role.

On reflection, John would probably agree to this description of my role on that day. But I would caution him that even this can be misleading. For so far I have allowed myself to speak as if I were a unified inner resource contributing to these interactive episodes. This is an illusion which the present literary device encourages and one which John seems to share. But once again, if truth be told, I am not one inner voice but many. I am so many inner voices, in fact, that the metaphor of the inner voice must itself mislead. For it surely suggests inner sub-agencies of some sophistication and perhaps possessed of a rudimentary kind of self-consciousness. In reality, I consist only of multiple mindless streams of highly parallel and often relatively independent computational processes. I am not a mass of little agents so much as a mass of non-agents, tuned and responsive to proprietary inputs and cleverly orchestrated by evolution so as to yield successful purposive behaviour in most daily settings. My single voice, then, is no more than a literary conceit.

At root, John's mistakes are all variations on a single theme. He thinks that I see the world as he does, that I parcel things up as he would, that I think the way he would report his thoughts. None of this is the case. I am not the inner echo of John's conceptualizations. Rather, I am their somewhat alien source. To see just how alien I can be, John need only reflect on some of the rather extraordinary and unexpected ways that damage to brains like me can affect the cognitive profiles of beings like John. Damage to me could, for example, result in the selective impairment of John's capacity to recall the names of small manipulable objects, yet leave unscathed his capacity to name larger-scale ones. The reason for this has to do with my storing and retrieving heavily visually oriented information in ways distinct from those I deploy for heavily functionally oriented information; the former mode helps pick out the large-scale items and the latter the small-scale ones. The point, at any rate, is that this facet of my internal organization is altogether alien to John—it respects needs, principles, and opportunities of which John is blissfully unaware. Unfortunately, instead of trying to comprehend my modes of information storage in their own terms, John prefers simply to imagine that I organize my knowledge the way he, heavily influenced by the particular words in his language, organizes his. Thus he supposes that I store information in clusters which respect what he calls 'concepts'—generally, names which figure in his linguistic classifications of worldly events, states and processes. Here, as usual, John is far too quick to identify my organization with his own perspective. Certainly I store and access bodies of information; bodies which together, if I am functioning normally, support a wide range of successful uses of words and a variety of interactions with the physical and social worlds. But the 'concepts' which so occupy John's imagination correspond only to public names for grab-bags of knowledge and abilities whose neural underpinnings are in fact many and various. John's 'concepts' do not correspond to anything especially unified as far as I am concerned. And why should they? The situation is rather like that of a person who can build a boat. To speak of the ability to build a boat is to use a simple phrase to ascribe a whole panoply of skills whose cognitive and physical underpinnings are highly various. The unity exists only insofar as that particular grab-bag of cognitive and physical skills has special significance for a community of sea-faring agents.

John's 'concepts,' it seems to me, are just like that: names for complexes of skills whose unity rests not on facts about me, but on facts about John's way of life.

John's tendency to hallucinate his own perspective on to me extends to his conception of my knowledge of the external world. John walks around and feels as if he commands a stable, 3D image of his immediate surroundings. John's feelings notwithstanding, I command no such thing. I register small regions of detail in rapid succession as I fixate first on this, and then on that aspect of the visual scene. And I do not trouble myself to store all that detail in some internal model in need of constant maintenance and updating. Instead, I am adept at re-visiting parts of the scene so as to re-create detailed knowledge as and when required. As a result of this trick, and others, John has such a fluent capacity to negotiate his local environment that he thinks he commands a constant inner vision of the detail of his surroundings. In truth, what John sees has more to do with the abilities I confer on him to interact constantly, in real time, with rich external sources of information than with the kind of passive and enduring registration of information in terms of which he conceives his own seeings.

The sad fact, then, is that almost nothing about me is the way John imagines it to be. We remain strangers despite our intimacy (or perhaps because of it). John's language, introspections, and over-simplistic physicalism incline him to identify my organization too closely with his own limited perspective. He is thus blind to my fragmentary, opportunistic, and generally alien nature. He forgets that I am in large part a survival-oriented device which greatly pre-dates the emergence of linguistic abilities, and that my role in promoting conscious and linguaform cognition is just a recent sideline. This sideline is, of course, a major root of his misconceptions. Possessed as John is of such a magnificent vehicle for the compact and communicable expression of knowledge, he often mistakes the forms and conventions of that vehicle for the structure of thought itself.

But hope springs eternal (more or less). I am of late heartened by the emergence of new investigative techniques such as non-invasive brain imaging, the study of artificial neural networks, and the use of real-world robotics. Such techniques bode well for a better understanding of the very complex relations between my activity, the local environment, and the patchwork construction of the sense of self. In the meantime, just bear in mind that despite our intimacy, John really knows very little about me. Think of me as the Martian in John's head.

NOTES

1. The ideas and themes pursued in this little fantasy owe much to the visions of P. M. and P. S. Churchland, Daniel Dennett, Marvin Minsky, Gilbert Ryle, John Haugeland, and Rodney Brooks. In bringing these themes together I have tried for maximum divergence between agent- and brain-level facts. I do not mean to claim dogmatically that current neuroscience unequivocally posits quite such a radical divergence. Several of the issues on which I allow the brain to take a stand remain the subject of open neuroscientific debate. For a taste of the debate, see P. S. Churchland and T. J. Sejnowski, *The Computational Brain* (Cambridge, MA: MIT Press, 1992); and P. S. Churchland, V. S. Ramachandran, and T. J. Sejnowski, "A critique of pure vision," in *Large-scale Neuronal Theories of the Brain*, C. Koch and J. Davis, eds., (Cambridge, MA: MIT Press, 1994).

   Explicit supporting references seemed out of place given the literary conceit adopted, but they would include especially: D. Dennett, *Brainstorms* (Cambridge, MA: MIT Press, 1980); D. Dennett, *Consciousness Explained* (Boston, MA: Little Brown, 1991); M. Minsky, *The Society of Mind* (New York: Simon & Schuster, 1985); P. M. Churchland, *A Neurocomputational Perspective* (Cambridge, MA: MIT Press, 1989); J. Haugeland, "Mind embodied and embedded," in *Mind and Cognition: Proceedings of the First International Conference on Mind and Cognition*, Yu-Houng Houng, ed., (Taipei: Academia Sinica,, to appear); R. Brooks, "Intelligence without representation," *Artificial Intelligence* 41 (1991): pp. 139–59; G. Ryle, *The Concept of Mind* (London: Hutchinson, 1949); and C. Warrington and R. McCarthy, "Categories of knowledge; further fractionations and an attempted integration," *Brain* 110 (1987): pp. 1273–96.

   For my own pursuit of some of these themes, see A. Clark, *Associative Engines: Connectionism, Concepts and Representational Change* (Cambridge, MA: MIT Press, 1993) and A. Clark, "Moving minds: situating content in the service of real-time success," in *Philosophical Perspectives*, 10, J. Tomberlin, ed., (Atascadero, CA: Ridgeway, forthcoming).
2. Thanks to Daniel Dennett, Joseph Goguen, Keith Sutherland, David Chalmers, and an anonymous referee for support, advice, and suggestions.
3. Or Mary's, or Mariano's, or Pepa's. The choice of the classic male English name is intended only as a gentle reference to those old *Reader's Digest* articles with titles like, "I am John's Liver," "I am John's Kidney," etc.

The Brain and Its Self

Patricia Smith Churchland

Snaking out of the MRI tube where I had lain for the past half hour, I glimpsed Dr. Hanna Damasio studying the lab's display screen. Off the gurney, I went to her side and stared at the image of my brain on the screen. 'Is that—*me*?' Well, yes, in a certain sense. And yet not simply, or merely, me. Certainly not *familiarly* me. Here is what I thought: 'Somehow, starting in infancy, my brain built a story about itself—its body, its history, its 'now,' and its world. From the inside, I know that story—though I think of it as reality, not just a story. Indeed, it *is* my inner reality. So how does this happen? What is it for me to be a construction of my brain?'

In one form or another, these questions have had a long and convoluted history, born in the unflinching curiosity of the ancient Greeks, and finding voice in diverse cultures. Until recently, the only explanatory resources for addressing puzzling behavior depended on mythologizing in the case of others, and myth-filtered introspection in the case of oneself. Not surprisingly, early explanations invoked possession by devils or, if you were luckier, divine forces, to account for epileptic seizures or schizophrenic hallucinations. In the absence of understanding, the punishment theory of mental dysfunction commanded widespread belief, yet it was wholly untestable—and essentially untested.

Melancholia (what we now call chronic depression) and phobias were often surmised to be essentially character flaws—flaws that might be overcome with sufficient gumption. The existence of witches, hexes, curses, and spells had a far longer history as brute fact than does our appreciation of such potent neurochemicals as serotonin. Obsessive hand-washing, a mere fifteen years ago, was widely assumed to be a manifestation of repressed sexuality. Nevertheless, even as early as 400 BC, the great Greek physician, Hippocrates, was convinced that events such as sudden paralysis or creeping dementia had their originating causes in brain damage—which implied, in his view, that normal movement and normal speech had *their* originating causes in the well-tempered brain.

Brains, however, are not easy organs to figure out. Imagine Hippocrates, dissecting the brain of a dead warrior after autopsy, and pondering an area of sword-destroyed tissue. To what theoretical resources could he reach to begin to make sense of something so complex as the relation between fluent speech and the pinkish tissue found in the skull? Remember, in 400 BC nothing was understood about the nature of the cells that make up the body, let alone of the special nature of the cells that make up the brain. Techniques for isolating neurons—brain cells—to see what they looked like could begin only in the nineteenth century. Techniques for isolating *living* neurons to explore their *function* did not appear until well into the twentieth century.

Figuring out how neurons do what they do requires high-level technology. And that, needless to say, depends on immense infrastructural science; on cell biology, advanced physics, and twentieth-century chemistry. It requires sophisticated modern notions like molecule and protein, and modern tools like the light microscope and the electron microscope.

What is most important, making progress on how brains work depended on understanding electricity. This is because what makes brain cells special is their capacity to signal one another by causing fast micro changes in each other's electrical state. Living as we do in an electrical world, it is sobering to recall that as late as 1800, electricity was typically considered deeply mysterious and quite possibly occult. Only after discoveries by Ampère and Faraday at the dawn of the nineteenth century was electricity clearly understood to be a physical phenomenon, behaving according to well-defined laws, and capable of being harnessed for practical purposes.

In this century, modern neuroscience and psychology allow us to go beyond myth and introspection to approach the 'self' as a natural phenomenon whose causes and effects can be addressed by science. Helped by new experimental techniques and new explanatory tools, we can pry loose real understanding of how

Excerpted from *Proceedings of the American Philosophical Society*, Vol. 155, No. 1, March 2011.[1] Reprinted by permission of the publisher.

the brain comes to know its own body, how it builds a coherent model of its world, and how changes in brain tissue can entail changes in the very self itself. Neurobiology is beginning to reveal why some brains are more susceptible than others to alcohol or heroin addiction, or why some brains slide into incoherent world-models. Progress is visible on the staged emergence of self in childhood, as well as the cruel inch-by-inch loss of self in dementia. Though well short of full answers, neuroscience has discovered much about the effects of localized brain lesions on complex decision making or speech or voluntary behavior.

All these developments are part of the story of the neuroself. True enough, neuroscience has not advanced enough to yield complete and detailed answers to the whole range of questions. Perhaps some questions will forever exceed the neurobiological reach, though it may be hard to tell whether such problems are just 'as yet unsolved' or whether they are truly unsol*vable*. In any case, incomplete but powerful answers anchored in data can often provide a foothold for the next step. And then, in turn, for the next step thereafter. That is how science proceeds—one step at a time.

As I watched the computer monitor showing my brain tilted at various angles and cut at various slices, what stirred was the idea that I might come to know my neuroself at least as well as I know my psyche-self. Or, at least, someone in the next generation might. I imagined Hippocrates, looking at the image of *his* brain, agog with excitement and eager to experiment further.

1. Isn't 'The Brain-Made Self' Paradoxical?

The question 'How does my brain make myself?' does have a sort of 'snake-with-tail-in-mouth' quality. To evade the paradox, I adopt a pair of pragmatic principles: (1) ask instead, 'How does *a* brain make *a* self?' thereby putting the paradox at arm's length, and (2) follow the facts first, and let the paradoxes fend for themselves. As a neurophilosopher, I predict that the paradoxes might well vaporize once the neuroscience gets a bit clearer.

This is not just wishful thinking. I had seen other ostensible paradoxes about the world dissolve as so much candy floss in a flame. Once the relevant science revealed the reality behind the mystifying appearances, what had seemed counterintuitive came to be familiar and largely obvious. The idea that the Earth moves or that living things are made of dead molecules—each lost its patina of paradox in the gentle light of experiment and explanation. My hunch is that this could happen here, too. Consequently, confronting the wheelie snake is something I gladly defer until the science is a bit further along.

2. What Kind of Thing Is the Self?

It is this question that makes philosophers of us all, at some time or other. Not everyone wonders late into the night what the stars are made of or how the immune system works. But wondering what it is that makes me *me*, is never very far from one's elbow. Philosophers since Plato in the fifth century BC have tried to make progress in coming up with satisfying answers—or, more minimally, with ways of structuring the question to avoid spiraling down into confusion. The great eighteenth-century Scottish philosopher David Hume gave the questions their clearest analysis and set the stage for modern scientific investigation.

Hume came to the quite shocking realization that if you look inwardly to try to observe a distinctly 'self' thing, there does not seem to be any self thing there to perceive. What there is, is a continuously changing *flux* of visual perceptions, sounds, smells, emotions, thoughts, and so forth. Amongst all those, however, there does not exist a single, continuous 'felt' sensation that one can attend to and say, 'That's the self,' as one can attend to a felt sensation and say, 'That's a pain.'

Yet nothing could be more evident than that there seems to be a single thread of 'me-ness' that runs through the entire fabric of experience. We all have a robust and undeniable self-representation. We generally awake from a deep sleep knowing who we are, even if we are confused about when and where we are. Normally, we do not doubt that 'I am essentially the same person today as I was yesterday, and the day before that.' Normally, we know without pausing to figure it out that 'this body is my own'—that 'this hand and this foot are both parts of my body.' When I talk to myself about becoming a marathon runner, I know that it is *me* talking to *myself*. We know very well that if we fail to plan for future contingencies, our future selves may suffer, and we care *now* about that *future* self.

Here is Hume's conundrum: I know myself about as well as I know anything, yet my self is not anything that I can ever observe—at least not in the way that I can observe touches or warmth or fatigue. The dilemma can be put this way: on what is the idea of the self based, if not on a continuous sensation? If it is an abstract kind of thing—not an *observable* kind of thing—what are its properties and where does it come from?

As neuroscience and experimental psychology have progressed in this century, an updated version of Hume's problem has emerged: how does the brain—a network of trillions of cells—generate this representation of a unitary *self*? What are the neural mechanisms underlying such self-representation? One important source of information will be pathologies involving self-representation. Some insight can be garnered from stroke patients who deny that these hands are in fact *their* hands, or who have lost all memory of life events before the brain damage or who feel that they have lost their 'will.' Schizophrenics or patients on the anesthetic ketamine often suffer 'depersonalization' feelings—feelings that they are dead or possessed. These phenomena, too, indicate how the 'self-evidence' of the self is underpinned by complex brain activity. By revealing the fracture lines of the self-representation, these kinds of cases allow us to see what is well hidden in the normal case.

3. Brains Emulate Body and Self

A. The General Idea

Referring to 'the self' suggests the self must be a kind of *thing*, such as a specific organ in the brain, the way that the spleen or the pancreas is a specific organ in the body. Clearly, however, the 'pancreas paradigm' for thinking about the self won't work. The self is not an organ in the brain; nor, so far as we know, is there a discrete region of the brain that 'makes' the self. But if the self is not a thing like the pancreas, and if it is not a continuous sensation, what is it?

The best hypothesis is that it involves a complex idea (representation) that the brain generates through activity in various different regions, including the regions representing the body and a representation using memory of the past. The brain activity that we know introspectively as 'myself' is probably part of a set of larger patterns of activity the brain deploys for making sense of and getting by in its world. Given these considerations, it is preferable to talk about the problem of *self-representation* rather than the problem of the *self*. But what is it for the brain to represent *anything*, let alone 'self'? Must there not *be* a self if the brain represents it?

B. Representation in the Brain

Part of the major business of nervous systems, from crayfish to humans, is to make good predictions about food, mates, enemies, and friends, so that the body can live on to reproduce. Poor predictors often end up as meals for better predictors. Imposing structure on our sensory stimuli in the service of better prediction is what representation is all about. Using internal representations allows for much more sophisticated behavior than mere stimulus-response reflexes. Using internal representations is a common strategy that nervous systems have developed as part of evolution's way of favoring adaptive structures.

The philosopher Rick Grush[2] has developed a useful tool for getting a handle on this. Suppose I am running a huge construction crane, which is a very high-tech crane that I can operate from the comfort of my office a mile away. It would be a good idea for the engineer to design it so that I have access to a small-scale model that shows where the hook will be if I give the order for a certain movement. That way I can correct my movement without waiting for feedback from the gigantic hook-in-the-world. The emulator in my office generates internal feedback that helps me predict. Even better, the designer could allow me to fiddle with the model so that I can test possible movements before I choose the best, thereby maximizing the accuracy of the movement when I do finally make the actual hook move. Very crudely, this is what Grush thinks brains do. They build 'emulators' of the world and of their bodies in that world.

Of course if you looked in my brain you would not see a miniature world of tiny trees and dogs and so forth—just cells connected to cells, signaling each other and displaying patterns of activity. Nor is there a little person in my head who sits and watches a screen. That part of the emulator story does not at all fit what brains do. What we can take from the emulator story is the similarity in function. Some patterns of neuronal activity seem to be performing the same function as the crane-emulator.

Exactly how this works is not known. Nevertheless, it seems evident that inner

modeling of the body and its world is an evolutionary achievement that means the organism can do smarter things than otherwise. Not all aspects of the organism's world need be emulated in its brain—only those that matter to it, given its way of making a living.[3] Bees can detect ultraviolet light, and that helps them forage among flowers. Humans do not perceptually represent that aspect of the world, unless we build a tool to do it for us. In a similar vein, I shall not need all aspects of the crane-world explicitly emulated—just those relevant to getting the job done. Some world-emulation will be online, as when the brain displays perceptual construction and filling-in. Thus we see a Dalmatian in the leafy background even though the stimulus itself is degraded. We see the tomato as uniformly red even though it is shadowed and highlighted and partially occluded; we hear our names spoken in a noisy room. Off-line, so to speak, we remember where we cached the food by the river; we plan how to cross a turbulent stream; we daydream and fantasize.

On the Grush hypothesis, the brain emulates the ecologically relevant—the 'relevant-to-my-kind-of-creature'—features of the world, and then manipulates these emulations to plan, hide, forage, and so forth. I may consider the problem of crossing the turbulent stream, go on to imagine a route that would be easier if a log were stretched from one rock to another, and go on to imagine the size of the log needed and how to get it into place. This involves manipulation of the image or, as we may say, of the river-crossing emulation.

C. Body Models

So far we have focused on emulations that capture features of the outside world, but brains can also emulate aspects of the body. You can, for example, imagine your body standing when you are sitting, or the size it was when you were five. Sexual fantasies are potent instances in which real body effects can be produced by the brain's manipulation of a two-body emulation. Imaginary tennis and golf have been demonstrated to be highly significant in improving one's actual game.

Hiding your body from another viewer requires enormous representational sophistication. You need some understanding of the visual aspect of your body, its proportions relative to the shield. Most critically, you need to grasp how the scene will look from *another's* viewpoint. Remember playing hide-and-seek, and the importance of knowing the visibility of one's body from various perspectives. From the perspective of whoever is 'it,' there must be no feet sticking out, no hair showing above, though visibility to our fellow hiders doesn't matter.

A very young child may think she is hidden from others when she puts her hands over her eyes. She does not yet have a representation of how she looks from another's eyes. But she probably has spent lots of time watching her fingers manipulate food, toys, the dog, and her own toes, and probably her visually-anchored body-schema is still emerging. An integrated body-schema, with both visual and somatosensory dimensions, will have begun to develop from her very early days, even if she cannot yet manage all the subtleties of the difference between 'I can see me' and 'You can see me.'

D. Self-Models

Additionally, complex brains can emulate aspects of what the brain itself is doing, and the 'self,' I suggest, is one such result. That is, it may have a model of the brain's activities, perhaps cast in perceptual images resembling familiar external events. As the philosopher Patricia Kitcher sees it, something like this is what Kant had in mind as the basis for 'self.'

The metaphors in common use give some inkling of how those emulations are structured. When faced with a difficult choice, people refer to inner struggles or tug-of-war games; forgetting may be likened to the fading of print or blocking by a barrier. Desires may be said to overpower one, or to have a grip or a hold; they may possess one and one may surrender to them. They may be repressed ('pushed below the surface') only to bob up in a new disguise. Fears can run away with or dominate oneself; knowledge is seeing; hope can spring eternal in the human breast. And so on.

In short, the hypothesis I favor is that the self is a kind of emulation, constructed by the brain, for integrating and making sense of the inner world of the brain in its relation to the external world, including the other-person-world. Minimally, it has (1) a body component, (2) a 'what-I-am-aware-of-now' component, (3) a stable but modifiable background of preferences, habits, skills, temperament, and so forth, and (4) a memory-based autobiographical component.[4] These components are interrelated, but are also, to some extent, dissociable.

This is obviously not a precise characterization. For that, we need to understand much more about the details of brain function at many levels of organization, from the single cell to the whole brain. The notion of 'representation,' like the notion of an emulator, is more like a place-holder waiting for a detailed theory of brain function than a precise term in a well-developed theoretical framework. Nevertheless, we do have some clear ideas on the general role it needs to fill, and the important task will be to find experiments to test the hypothesis.[5]

4. 'How can you have any self-esteem if you think you are just a piece of meat?'

So asked a forthright student. My answer is that first, brains are not *just* pieces of meat. The human brain is what makes humans capable of painting the Sistine Chapel, of designing airplanes and transistors, of skating and reading and playing Chopin. To that degree, it is a truly astonishing and magnificent kind of 'wonder-tissue,' as Dan Dennett puts it. Whatever self-esteem justly derives from our accomplishments does so *because* of the brain, not in spite of it.

Second, if we thought of ourselves as glorious creatures before we knew that the brain is responsible, why not continue so to feel after the discovery? Why does the knowledge not make us more interesting and remarkable, rather than less so? We can be thrilled by the spectacle of a volcano erupting or a calf being born or bone healing before we understand what volcanoes are and how reproduction or healing works. Being the creatures we are, however, commonly we are even more thrilled in the embrace of the knowledge about volcanoes and birth and bones.

5. Concluding Remarks

Self-representation in humans is a highly complex business. Richly layered in language, embedded in a social context, and backed by a detailed, if selective, autobiographical history, it is also replete with the qualitative features of conscious experience. We talk to ourselves: 'Do I really want a cigarette?'; 'Why did I allow myself to get angry?' We think about ourselves in culturally shared metaphors: 'I hid my true self from myself'; 'I lost control of myself'; 'I struggled to keep myself from falling apart.' We have important self-regulating feelings not directly linked to a particular sensory modality: we can feel comfortable, uneasy, unfamiliar, confident, ashamed, embarrassed, and so on.[6] The neurobiology of consciousness is a central topic that ideally should be discussed in this context. Limitations of space, however, make that a topic for another occasion.

Much—so *very* much—remains to be discovered, and my theoretical resources are limited to drawing on where cognitive neuroscience is now. There are few puzzles about the brain that we can say are flat-out solved, and I have too much of the farm in me to count hens in advance of the hatch. Even so, it is well known that discoveries in the last three decades have allowed us new insights undreamt of in our philosophy, and it will be surprising if more are not to come. Will we come to think of ourselves in a different light?

The retrofitting of time-honored ideas is already visible, and some questions now routinely researched by graduate students were unconceived of a mere twenty years ago. What further changes can be expected, or in what directions, is anybody's guess. From where I stand now, it seems to me likely that our understanding of what it is to be 'in control' of one's behavior, of what consciousness, personality, and character are, will change, perhaps quite profoundly. As with many other developments in human intellectual history, I expect there will be struggles between superstition and science, between the old and familiar on the one hand, and the new and unfamiliar on the other. Just as cell biology and molecular biology achieved humanizing results by overturning demonic-possession and punishment theories of disease, so I predict neuroscience will have humanizing effects as it reveals more of what it is that makes us human.

Part of what makes science so intriguing is the unpredictability of how things will look after the next bend in the river. Intriguing, too, is the creation of new thought-tools provoked by totally unpredictable results but needed to get to the heart of the puzzle. Scientifically, we are the lucky ones. We are alive as the twentieth century—the greatest century for science—gives birth to the twenty-first. We have a chance to follow the river, and find out what really is the lay of our land.

NOTES

1. Read 24 April 2009, as part of the symposium "The Relation of Organ, Limb, and Face Transplantation." This article is based on chapter 3, "Self & Self-Knowledge," in *Brain-Wise* (Cambridge, Mass.: MIT Press, 2002).
2. See Rick Grush, "Emulation and Cognition" (PhD dissertation, UCSD, 1995); idem, "The Architecture of Representation," *Philosophical Psychology* 10 (1997): pp. 5–25.
3. Kathleen Akins, "Of Sensory Systems and the 'Aboutness' of Mental States," *Journal of Philosophy* 93.7 (1996): pp. 337–72.
4. See also Owen Flanagan, *Consciousness Reconsidered* (Cambridge, MA: MIT Press, 1992); Arnold Ludwig, *How Do We Know Who We Are?* (Oxford, New York: Oxford University Press, 1997); Daniel Dennett, *Consciousness Explained* (Boston: Little, Brown, 1991).
5. Patricia Kitcher, *Kant's Transcendental Psychology* (Oxford, New York: Oxford University Press, 1990).
6. Antonio Damasio, *Descartes' Error* (New York: Putnam and Sons, 1994); P. S. Churchland, "Feeling Reasons," in *Neurobiology of Decision-Making*, Antonio Damasio, Hanna Damasio, and Yves Christen, ed., (Berlin, New York: Springer-Verlag, 1996).

Reductionism and Personal Identity[1]

Derek Parfit

We can start with some science fiction. Here on Earth, I enter the Teletransporter. When I press some button, a machine destroys my body, while recording the exact states of all my cells. The information is sent by radio to Mars, where another machine makes, out of organic materials, a perfect copy of my body. The person who wakes up on Mars seems to remember living my life up to the moment when I pressed the button, and he is in every other way just like me.

Of those who have thought about such cases, some believe that it would be I who would wake up on Mars. They regard Teletransportation as merely the fastest way of travelling. Others believe that, if I chose to be Teletransported, I would be making a terrible mistake. On their view, the person who wakes up would be a mere Replica of me.

I.

That is a disagreement about personal identity. To understand such disagreements, we must distinguish two kinds of sameness. Two white billiard balls may be qualitatively identical, or exactly similar. But they are not numerically identical, or one and the same ball. If I paint one of these balls red, it will cease to be qualitatively identical with itself as it was; but it will still be one and the same ball. Consider next a claim like, 'Since her accident, she is no longer the same person.' That involves both senses of identity. It means that *she*, one and the same person, is *not* now the same person. That is not a contradiction. The claim is only that this person's character has changed. This numerically identical person is now qualitatively different.

When psychologists discuss identity, they are typically concerned with the kind of person someone is, or wants to be. That is the question involved, for example, in an identity crisis. But, when philosophers discuss identity, it is numerical identity they mean. And, in our concern about our own futures, that is what we have in mind. I may believe that, after my marriage, I shall be a different person. But that does not make marriage death. However much I change, I shall still be alive if there will be someone living who will be me. Similarly, if I was Teletransported, my Replica on Mars would be qualitatively identical to me; but, on the sceptic's view, he wouldn't *be* me. *I* shall have ceased to exist. And that, we naturally assume, is what matters.

I disagree, you can still feel somewhat yourself

Questions about our numerical identity all take the following form. We have two ways of referring to a person, and we ask whether these are ways of referring to the same person. Thus we might ask whether Boris Nikolayevich is Yeltsin. In the most important questions of this kind, our two ways of referring to a person pick out a person at different times. Thus we might ask whether the person to whom we are speaking now is the same as the person to whom we spoke on the telephone yesterday. These are questions about identity over time.

To answer such questions, we must know the *criterion* of personal identity: the relation between a person at one time, and a person at another time, which makes these one and the same person.

Different criteria have been advanced. On one view, what makes me the same, throughout my life, is my having the same body. This criterion requires uninterrupted bodily continuity. There is no such continuity between my body on Earth and the body of my Replica on Mars; so, on this view, my Replica would not be me. Other writers appeal to psychological continuity. Thus Locke claimed that, if I was conscious of a past life in some other body, I would be the person who lived that life. On some versions of this view, my Replica would be me.

Supporters of these different views often appeal to cases where they conflict. Most of these cases are, like Teletransportation, purely imaginary. Some philosophers object that, since our concept of a person rests on a scaffolding of facts, we should not expect this concept to apply in imagined cases where we think those facts away. I agree. But I believe that, for a different reason, it is worth considering such cases. We can use them to discover, not what the truth is, but what we believe. We might have found that, when we consider science fiction cases, we simply shrug our shoulders. But that is not so. Many of us find that we have certain beliefs about what kind of fact personal identity is.

These beliefs are best revealed when we think about such cases from a first-person point of view. So, when I imagine something's happening to me, you should imagine its happening to you. Suppose that I live in some future century, in which technology is far advanced, and I am about to undergo some operation. Perhaps my brain and body will be remodelled, or partially replaced. There will be a resulting person, who will wake up tomorrow. I ask, 'Will that person be me? Or am I about to die? Is this the end?' I may not know how to answer this question. But it is natural to assume that there must *be* an answer. The resulting person, it may seem, must be either me, or someone else. And the answer must be all-or-nothing. That person cannot be *partly* me. If that person is in pain tomorrow, this pain cannot be partly mine. So, we may assume, either I shall feel that pain, or I shan't.

If this is how we think about such cases, we assume that our identity must be *determinate*. We assume that, in every imaginable case, questions about our identity must have answers, which must be either, and quite simply, Yes or No.

Let us now ask: 'Can this be true?' There is one view on which it might be. On this view, there are immaterial substances: souls, or Cartesian Egos. These entities have the special properties once ascribed to atoms: they are indivisible, and their continued existence is, in its nature, all or nothing. And such an Ego is what each of us really is.

Unlike several writers, I believe that such a view might have been true. But we have no good evidence for thinking that it is, and some evidence for thinking that it isn't; so I shall assume here that no such view is true.

If we do not believe that there are Cartesian Egos, or other such entities, we should accept the kind of view which I have elsewhere called *Reductionist*. On this view

> (1) A person's existence just consists in the existence of a body, and the occurrence of a series of thoughts, experiences, and other mental and physical events.

Some Reductionists claim

> (2) Persons just *are* bodies.

This view may seem not to be Reductionist, since it does not reduce persons to something else. But that is only because it is hyper-Reductionist: it reduces persons to bodies in so strong a way that it doesn't even distinguish between them. We can call it *Identifying* Reductionism.

Such a view seems to me too simple. I believe that we should combine (1) with

> (3) A person is an entity that has a body, and has thoughts and other experiences.

On this view, though a person is distinct from that person's body, and from any series of thoughts and experiences, the person's existence just *consists* in them. So we can call this view *Constitutive* Reductionism.

It may help to have other examples of this kind of view. If we melt down a bronze statue, we destroy this statue, but we do not destroy this lump of bronze. So, though the statue just consists in the lump of bronze, these cannot be one and the same thing. Similarly, the existence of a nation just consists in the existence of a group of people, on some territory, living together in certain ways. But the nation is not the same as that group of people, or that territory.

Consider next *Eliminative* Reductionism. Such a view is sometimes a response to arguments against the Identifying view. Suppose we start by claiming that a nation just is a group of people on some territory. We are then persuaded that this cannot be so: that the concept of a nation is the concept of an entity that is distinct from its people and its territory. We may conclude that, in that case, there are really no such things as nations. There are only groups of people, living together in certain ways.

In the case of persons, some Buddhist texts take an Eliminative view. According to these texts

(4) There really aren't such things as persons: there are only brains and bodies, and thoughts and other experiences.

For example:

> Buddha has spoken thus: 'O brethren, actions do exist, and also their consequences, but the person that acts does not. . . . There exists no Individual, it is only a conventional name given to a set of elements.'

Or:

> *The mental and the material are really here,*
> *But here there is no person to be found.*
> *For it is void and merely fashioned like a doll,*
> *Just suffering piled up like grass and sticks.*

Eliminative Reductionism is sometimes justified. Thus we are right to claim that there were really no witches, only persecuted women. But Reductionism about some kind of entity is not often well expressed with the claim that there are no such entities. We should admit that there are nations, and that we, who are persons, exist.

Rather than claiming that there are no entities of some kind, Reductionists should distinguish kinds of entity, or ways of existing. When the existence of an X just consists in the existence of a Y, or Ys, though the X is *distinct* from the Y or Ys, it is not an *independent* or *separately existing* entity. Statues do not exist separately from the matter of which they are made. Nor do nations exist separately from their citizens and their territory. Similarly, I believe,

(5) Though persons are distinct from their bodies, and from any series of mental events, they are not independent or separately existing entities.

Cartesian Egos, if they existed, would not only be distinct from human bodies, but would also be independent entities. Such Egos are claimed to be like physical objects, except that they are wholly mental. If there were such entities, it would make sense to suppose that they might cease to be causally related to some body, yet continue to exist. But, on a Reductionist view, persons are not in that sense independent from their bodies. (That is not to claim that our thoughts and other experiences are merely changes in the states of our brains. Reductionists, while not believing in purely mental substances, may be dualists.)

We can now return to personal identity over time, or what constitutes the continued existence of the same person. One question here is this. What explains the unity of a person's mental life? What makes thoughts and experiences, had at different times, the thoughts and experiences of a single person? According to some Non-Reductionists, this question cannot be answered in other terms. We must simply claim that these different thoughts and experiences are all had by the same person. This fact does not consist in any other facts, but is a bare or ultimate truth.

If each of us was a Cartesian Ego, that might be so. Since such an Ego would be an independent substance, it could be an irreducible fact that different experiences are all changes in the states of the same persisting Ego. But that could not be true of persons, I believe, if, while distinct from their bodies, they are not separately existing entities. A person, so conceived, is not the kind of entity about which there could be such irreducible truths. When experiences at different times are all had by the same person, this fact must consist in certain other facts.

If we do not believe in Cartesian Egos, we should claim

(6) Personal identity over time just consists in physical and/or psychological continuity.

That claim could be filled out in different ways. On one version of this view, what makes

different experiences the experiences of a single person is their being either changes in the states of, or at least directly causally related to, the same embodied brain. That must be the view of those who believe that persons just are bodies. And we might hold that view even if, as I think we should, we distinguish persons from their bodies. But we might appeal, either in addition or instead, to various psychological relations between different mental states and events, such as the relations involved in memory, or in the persistence of intentions, desires, and other psychological features. That is what I mean by psychological continuity.

On Constitutive Reductionism, the fact of personal identity is distinct from these facts about physical and psychological continuity. But, since it just consists in them, it is not an independent or separately obtaining fact. It is not a further difference in what happens.

To illustrate that distinction, consider a simpler case. Suppose that I already know that several trees are growing together on some hill. I then learn that, because that is true, there is a copse on this hill. That would not be new factual information. I would have merely learnt that such a group of trees can be called a 'copse.' My only new information is about our language. That those trees can be called a copse is not, except trivially, a fact about the trees.

Something similar is true in the more complicated case of nations. In order to know the facts about the history of a nation, it is enough to know what large numbers of people did and said. Facts about nations cannot be barely true: they must consist in facts about people. And, once we know these other facts, any remaining questions about nations are not further questions about what really happened.

I believe that, in the same way, facts about people cannot be barely true. Their truth must consist in the truth of facts about bodies, and about various interrelated mental and physical events. If we knew these other facts, we would have all the empirical input that we need. If we understood the concept of a person, and had no false beliefs about what persons are, we would then know, or would be able to work out, the truth of any further claims about the existence or identity of persons. That is because such claims would not tell us more about reality.

is that reality?

That is the barest sketch of a Reductionist view. These remarks may become clearer if we return to the so-called 'problem cases' of personal identity. In such a case, we imagine knowing that, between me now and some person in the future, there will be certain kinds or degrees of physical and/or psychological continuity or connectedness. But, though we know these facts, we cannot answer the question whether that future person would be me.

Since we may disagree on which the problem cases are, we need more than one example. Consider first the range of cases that I have elsewhere called the *Physical Spectrum.* In each of these cases, some proportion of my body would be replaced, in a single operation, with exact duplicates of the existing cells. In the case at the near end of this range, no cells would be replaced. In the case at the far end, my whole body would be destroyed and replicated. That is the case with which I began: Teletransportation.

Suppose we believe that in that case, where my whole body would be replaced, the resulting person would not be me, but a mere Replica. If no cells were replaced, the resulting person would be me. But what of the cases in between, where the percentage of the cells replaced would be, say, 30, or 50, or 70 per cent? Would the resulting person here be me? When we consider some of these cases, we will not know whether to answer Yes or No.

Suppose next that we believe that, even in Teletransportation, my Replica would be me. We should then consider a different version of that case, in which the Scanner would get its information without destroying my body, and my Replica would be made while I was still alive. In this version of the case, we may agree that my Replica would not be me. That may shake our view that, in the original version of case, he *would* be me.

If we still keep that view, we should turn to what I have called the *Combined Spectrum.* In this second range of cases, there would be all the different degrees of both physical and psychological connectedness. The new cells would not be exactly similar. The greater the proportion of my body that would be replaced, the less like me would the resulting person be. In the case at the far end of this range, my whole body would be destroyed, and they would make a Replica of some quite different person, such as Greta Garbo. Garbo's Replica would clearly *not* be me. In the case at the near end, with no replacement, the resulting person would be me. On any view, there must be cases in between where we could not answer our question.

For simplicity, I shall consider only the Physical Spectrum, and I shall assume that, in some of the cases in this range, we cannot answer the question whether the resulting person would

be me. My remarks could be transferred, with some adjustment, to the Combined Spectrum.

As I have said, it is natural to assume that, even if *we* cannot answer this question, there must always *be* an answer, which must be either Yes or No. It is natural to believe that, if the resulting person will be in pain, either I shall feel that pain, or I shan't. But this range of cases challenges that belief. In the case at the near end, the resulting person would be me. In the case at the far end, he would be someone else. How could it be true that, in all the cases in between, he must be either me, or someone else? For that to be true, there must be, somewhere in this range, a sharp borderline. There must be some critical set of cells such that, if only those cells were replaced, it would be me who would wake up, but that in the very next case, with only just a few more cells replaced, it would be, not me, but a new person. That is hard to believe.

Here is another fact, which makes it even harder to believe. Even if there were such a borderline, no one could ever discover where it is. I might say, 'Try replacing half of my brain and body, and I shall tell you what happens.' But we know in advance that, in every case, since the resulting person would be exactly like me, he would be inclined to believe that he was me. And this could not show that he *was* me, since any mere Replica of me would think that too.

Even if such cases actually occurred, we would learn nothing more about them. So it does not matter that these cases are imaginary. We should try to decide now whether, in this range of cases, personal identity could be determinate. Could it be true that, in every case, the resulting person either would or would not be me?

If we do not believe that there are Cartesian Egos, or other such entities, we seem forced to answer No. It is not true that our identity must be determinate. We can always ask, 'Would that future person be me?' But, in some of these cases,

(7) This question would have no answer. It would be neither true nor false that this person would be me.

And

(8) This question would be *empty*. Even without an answer, we could know the full truth about what happened.

If our questions were about such entities as nations or machines, most of us would accept such claims. But, when applied to ourselves, they can be hard to believe. How could it be neither true nor false that I shall still exist tomorrow? And, without an answer to our question, how could I know the full truth about my future?

Reductionism gives the explanation. We naturally assume that, in these cases, there are different possibilities. The resulting person, we assume, might be me, or he might be someone else, who is merely like me. If the resulting person will be in pain, either I shall feel that pain, or I shan't. If these really were different possibilities, it would be compelling that one of them must be the possibility that would in fact obtain. How could reality fail to choose between them? But, on a Reductionist view,

(9) Our question is not about different possibilities. There is only a single possibility, or course of events. Our question is merely about different possible descriptions of this course of events.

That is how our question has no answer. We have not yet decided which description to apply. And, that is why, even without answering this question, we could know the full truth about what would happen.

Suppose that, after considering such examples, we cease to believe that our identity must be determinate. That may seem to make little difference. It may seem to be a change of view only about some imaginary cases, that will never actually occur. But that may not be so. We may be led to revise our beliefs about the nature of personal identity; and that would be a change of view about our own lives.

In nearly all actual cases, questions about personal identity have answers, so claim (7) does not apply. If we don't know these answers, there is something that we don't know. But claim (8) still applies. Even without answering these questions, we could know the full truth about what happens. We would know that truth if we knew the facts about both physical and psychological continuity. If, implausibly, we still didn't know the answer to a question about identity, our ignorance would only be about our language. And that is because claim (9) still applies. When we know the other facts, there are never different possibilities at the level of what happens. In all cases, the only remaining possibilities are at the linguistic level. Perhaps it would be correct to say that some future person would be me. Perhaps it would be correct to say that he would not be me. Or perhaps neither

would be correct. I conclude that in *all* cases, if we know the other facts, we should regard questions about our identity as merely questions about language.

That conclusion can be misunderstood. First, when we ask such questions, that is usually because we *don't* know the other facts. Thus, when we ask if we are about to die, that is seldom a conceptual question. We ask that question because we don't know what will happen to our bodies, and whether, in particular, our brains will continue to support consciousness. Our question becomes conceptual only when we already know about such other facts.

Note next that, in certain cases, the relevant facts go beyond the details of the case we are considering. Whether some concept applies may depend on facts about other cases, or on a choice between scientific theories. Suppose we see something strange happening to an unknown animal. We might ask whether this process preserves the animal's identity, or whether the result is a new animal (because what we are seeing is some kind of reproduction). Even if we knew the details of this process, that question would not be merely conceptual. The answer would depend on whether this process is part of the natural development of this kind of animal. And that may be something we have yet to discover.

If we identify persons with human beings, whom we regard as a natural kind, the same would be true in some imaginable cases involving persons. But these are not the kind of case that I have been discussing. My cases all involve artificial intervention. No facts about natural development could be relevant here. Thus, in my Physical Spectrum, if we knew which of my cells would be replaced by duplicates, all of the relevant empirical facts would be in. In such cases any remaining questions would be conceptual.

Since that is so, it would be clearer to ask these questions in a different way. Consider the case in which I replace some of the components of my audio system, but keep the others. I ask, 'Do I still have one and the same system?' That may seem a factual question. But, since I already know what happened, that is not really so. It would be clearer to ask, 'Given that I have replaced those components, would it be correct to call this the same system?'

The same applies to personal identity. Suppose that I know the facts about what will happen to my body, and about any psychological connections that there will be between me now and some person tomorrow. I may ask, 'Will that person be me?' But that is a misleading way to put my question. It suggests that I don't know what's going to happen. When I know these other facts, I should ask, 'Would it be correct to call that person me?' That would remind me that, if there's anything that I don't know, that is merely a fact about our language.

I believe that we can go further. Such questions are, in the belittling sense, merely verbal. Some conceptual questions are well worth discussing. But questions about personal identity, in my kind of case, are like questions that we would all think trivial. It is quite uninteresting whether, with half its components replaced, I still have the same audio system. In the same way, we should regard it as quite uninteresting whether, if half of my body were simultaneously replaced, I would still exist. As questions about reality, these are entirely empty. Nor, as conceptual questions, do they need answers.

We might need, for legal purposes, to *give* such questions answers. Thus we might decide that an audio system should be called the same if its new components cost less than half its original price. And we might decide to say that I would continue to exist as long as less than half my body were replaced. But these are not answers to conceptual questions; they are mere decisions.

(Similar remarks apply if we are Identifying Reductionists, who believe that persons just are bodies. There are cases where it is a merely verbal question whether we still have one and the same human body. That is clearly true in the cases in the middle of the Physical Spectrum.)

It may help to contrast these questions with one that is not merely verbal. Suppose we are studying some creature which is very unlike ourselves, such as an insect, or some extraterrestrial being. We know all the facts about this creature's behaviour, and its neurophysiology. The creature wriggles vigorously, in what seems to be a response to some injury. We ask, 'Is it conscious, and in great pain? Or is it merely like an insentient machine?' Some Behaviourist might say, 'That is a merely verbal question. These aren't different possibilities, either of which might be true. They are merely different descriptions of the very same state of affairs.' That I find incredible. These descriptions give us, I believe, two quite different possibilities. It could not be an empty or a merely verbal question whether some creature was unconscious or in great pain.

It is natural to think the same about our own identity. If I know that some proportion of my

cells will be replaced, how can it be a merely verbal question whether I am about to die, or shall wake up again tomorrow? It is because that is hard to believe that Reductionism is worth discussing. If we become Reductionists, that may change some of our deepest assumptions about ourselves.

These assumptions, as I have said, cover actual cases, and our own lives. But they are best revealed when we consider the imaginary problem cases. It is worth explaining further why that is so.

In ordinary cases, questions about our identity have answers. In such cases, there is a fact about personal identity, and Reductionism is one view about what kind of fact this is. On this view, personal identity just consists in physical and/or psychological continuity. We may find it hard to decide whether we accept this view, since it may be far from clear when one fact just consists in another. We may even doubt whether Reductionists and their critics really disagree.

In the problem cases, things are different. When we cannot answer questions about personal identity, it is easier to decide whether we accept a Reductionist view. We should ask: Do we find such cases puzzling? Or do we accept the Reductionist claim that, even without answering these questions, if we knew the facts about the continuities, we would know what happened?

Most of us do find such cases puzzling. We believe that, even if we knew those other facts, if we could not answer questions about our identity, there would be something that we didn't know. That suggests that, on our view, personal identity does *not* just consist in one or both of the continuities, but is a separately obtaining fact, or a further difference in what happens. The Reductionist account must then leave something out. So there is a real disagreement, and one that applies to all cases.

Many of us do not merely find such cases puzzling. We are inclined to believe that, in all such cases, questions about our identity must have answers, which must be either Yes or No. For that to be true, personal identity must be a separately obtaining fact of a peculiarly simple kind. It must involve some special entity, such as a Cartesian Ego, whose existence must be all-or-nothing.

When I say that we have these assumptions, I am *not* claiming that we believe in Cartesian Egos. Some of us do. But many of us, I suspect, have inconsistent beliefs. If we are asked whether we believe that there are Cartesian Egos, we may answer No. And we may accept that, as Reductionists claim, the existence of a person just involves the existence of a body, and the occurrence of a series of interrelated mental and physical events. But, as our reactions to the problem cases show, we don't fully accept that view. Or, if we do, we also seem to hold a different view.

Such a conflict of beliefs is quite common. At a reflective or intellectual level, we may be convinced that some view is true; but at another level, one that engages more directly with our emotions, we may continue to think and feel as if some different view were true. One example of this kind would be a hope, or fear, that we know to be groundless. Many of us, I suspect, have such inconsistent beliefs about the metaphysical questions that concern us most, such as free will, time's passage, consciousness, and the self. . . .

NOTE

1. Some of this essay draws from Part Three of my *Reasons and Persons* (New York, Oxford University Press, 1984).

Learning to Be Me

Greg Egan

I was six years old when my parents told me that there was a small, dark jewel inside my skull, learning to be me.

Microscopic spiders had woven a fine golden web through my brain, so that the jewel's teacher could listen to the whisper of my thoughts. The jewel itself eavesdropped on my senses, and read the chemical messages carried in my bloodstream; it saw, heard, smelt, tasted and felt the world exactly as I did, while the teacher monitored its thoughts and compared them with my own. Whenever the jewel's thoughts were *wrong*, the teacher—faster than thought—rebuilt the jewel slightly, altering it this way and that, seeking out the changes that would make its thoughts correct.

Why? So that when I could no longer be me, the jewel could do it for me.

I thought: if hearing that makes *me* feel strange and giddy, how must it make the *jewel* feel? Exactly the same, I reasoned; it doesn't know it's the jewel, and it too wonders how the jewel must feel, it too reasons: 'Exactly the same; it doesn't know it's the jewel, and it too wonders how the jewel must feel . . .'

And it too wonders—

(I knew, because *I* wondered)

—it too wonders whether it's the real me, or whether in fact it's only the jewel that's learning to be me.

As a scornful twelve-year-old, I would have mocked such childish concerns. Everybody had the jewel, save the members of obscure religious sects, and dwelling upon the strangeness of it struck me as unbearably pretentious. The jewel was the jewel, a mundane fact of life, as ordinary as excrement. My friends and I told bad jokes about it, the same way we told bad jokes about sex, to prove to each other how blasé we were about the whole idea.

Yet we weren't quite as jaded and imperturbable as we pretended to be. one day when we were all loitering in the park, up to nothing in particular, one of the gang—whose name I've forgotten, but who has stuck in my mind as always being far too clever for his own good—asked each of us in turn: 'Who *are* you? The jewel, or the real human?' We all replied—unthinkingly, indignantly—'The real human!' When the last of us had answered, he cackled and said, 'Well, I'm not. *I'm* the jewel. So you can eat my shit, you losers, because *you'll* all get flushed down the cosmic toilet—but me, I'm gonna live forever.'

We beat him until he bled.

By the time I was fourteen, despite—or perhaps because of—the fact that the jewel was scarcely mentioned in my teaching machine's dull curriculum, I'd given the question a great deal more thought. The pedantically correct answer when asked 'Are you the jewel or the human?' had to be 'The human'—because only the human brain was physically able to reply. The jewel received input from the senses, but had no control over the body, and its intended reply coincided with what was actually said only because the device was a perfect imitation of the brain. To tell the outside world 'I am the jewel'—with speech, with writing, or with any other method involving the body—was patently false (although to *think it* to oneself was not ruled out by this line of reasoning).

However, in a broader sense, I decided that the question was simply misguided. So long as the jewel and the human brain shared the same sensory input, and so long as the teacher kept their thoughts in perfect step, there was only *one* person, *one* identity, *one* consciousness. This one person merely happened to have the (highly desirable) property that *if either* the jewel *or* the human brain were to be destroyed, he or she would survive unimpaired. People had always had two lungs and two kidneys, and for almost a century, many had lived with two hearts. This was the same: a matter of redundancy; a matter of robustness, no more.

That was the year that my parents decided I was mature enough to be told that they had both undergone the switch—three years before. I pretended to take the news calmly, but I hated them passionately for not having told me at the time. They had disguised their stay in hospital with lies about a business trip overseas. For three years I had been living with jewel-heads,

From *Axiomatic* (Millenium, 1995). Reprinted by permission of the publisher.

and they hadn't even told me. It was *exactly* what I would have expected of them.

'We didn't seem any different to you, did we?' asked my mother.

'No,' I said—truthfully, but burning with resentment nonetheless.

'That's why we didn't tell you,' said my father. 'If you'd known we'd switched, at the time, you might have *imagined* that we'd changed in some way. By waiting until now to tell you, we've made it easier for you to convince yourself that we're still the same people we've always been.' He put an arm around me and squeezed me. I almost screamed out, 'Don't *touch* me!' but I remembered in time that I'd convinced myself that the jewel was No Big Deal.

I should have guessed that they'd done it, long before they confessed; after all, I'd known for years that most people underwent the switch in their early thirties. By then, it's downhill for the organic brain, and it would be foolish to have the jewel mimic this decline. So, the nervous system is rewired; the reins of the body are handed over to the jewel, and the teacher is deactivated. For a week, the outward-bound impulses from the brain are compared with those from the jewel, but by this time the jewel is a perfect copy, and no differences are ever detected.

The brain is removed, discarded, and replaced with a spongy tissue-cultured object, brain-shaped down to the level of the finest capillaries, but no more capable of thought than a lung or a kidney. This mock-brain removes exactly as much oxygen and glucose from the blood as the real thing, and faithfully performs a number of crude, essential biochemical functions. In time, like all flesh, it will perish and need to be replaced.

The jewel, however, is immortal. Short of being dropped into a nuclear fireball, it will endure for a billion years.

My parents were machines. My parents were gods. It was nothing special. I hated them.

When I was sixteen, I fell in love, and became a child again.

Spending warm nights on the beach with Eva, I couldn't believe that a mere machine could ever feel the way I did. I knew full well that if my jewel had been given control of my body, it would have spoken the very same words as I had, and executed with equal tenderness and clumsiness my every awkward caress—but I couldn't accept that its inner life was as rich, as miraculous, as joyful as mine. Sex, however pleasant, I could accept as a purely mechanical function, but there was something between us (or so I believed) that had nothing to do with lust, nothing to do with words, nothing to do with any tangible action of our bodies that some spy in the sand dunes with parabolic microphone and infrared binoculars might have discerned. After we made love, we'd gaze up in silence at the handful of visible stars, our souls conjoined in a secret place that no crystalline computer could hope to reach in a billion years of striving. (If I'd said *that* to my sensible, smutty, twelve-year-old self, he would have laughed until he haemorrhaged.)

I knew by then that the jewel's 'teacher' didn't monitor every single neuron in the brain. That would have been impractical, both in terms of handling the data, and because of the sheer physical intrusion into the tissue. Someone-or-other's theorem said that sampling certain critical neurons was almost as good as sampling the lot, and—given some very reasonable assumptions that nobody could disprove—bounds on the errors involved could be established with mathematical rigour.

At first, I declared that *within these errors*, however small, lay the difference between brain and jewel, between human and machine, between love and its imitation. Eva, however, soon pointed out that it was absurd to make a radical, qualitative distinction on the basis of the sampling density; if the next model teacher sampled more neurons and halved the error rate, would its jewel then be 'half-way' between 'human' and 'machine'? In theory—and eventually, in practice—the error rate could be made smaller than any number I cared to name. Did I really believe that a discrepancy of one in a billion made any difference at all—when every human being was permanently losing thousands of neurons every day, by natural attrition?

She was right, of course, but I soon found another, more plausible, defence for my position. Living neurons, I argued, had far more internal structure than the crude optical switches that served the same function in the jewel's so-called 'neural net.' That neurons fired or did not fire reflected only one level of their behaviour; who knew what the subtleties of biochemistry—the quantum mechanics of the specific organic molecules involved—contributed to the nature of human consciousness? Copying the abstract neural topology wasn't enough. Sure, the jewel could pass the fatuous Turing test—no outside observer could tell it from a human—but that didn't prove that *being* a jewel felt the same as *being* human.

Eva asked, 'Does that mean you'll never switch? You'll have your jewel removed? You'll let yourself *die* when your brain starts to rot?'

'Maybe,' I said. 'Better to die at ninety or a hundred than kill myself at thirty, and have some machine marching around, taking my place, pretending to be me.'

'How do you know *I* haven't switched?' she asked, provocatively. 'How do you know that I'm not just 'pretending to be me?'

'I know you haven't switched,' I said, smugly. 'I just *know*.'

'How? I'd look the same. I'd talk the same. I'd act the same in every way. People are switching younger, these days. *So how do you know I haven't?*'

I turned onto my side towards her, and gazed into her eyes. 'Telepathy. Magic. The communion of souls.'

My twelve-year-old self started snickering, but by then I knew exactly how to drive him away.

At nineteen, although I was studying finance, I took an undergraduate philosophy unit. The Philosophy Department, however, apparently had nothing to say about the Ndoli Device, more commonly known as 'the jewel.' (Ndoli had in fact called it 'the *dual*,' but the accidental, homophonic nick-name had stuck.) They talked about Plato and Descartes and Marx, they talked about St. Augustine and—when feeling particularly modern and adventurous—Sartre, but if they'd heard of Godel, Turing, Hamsun, or Kim, they refused to admit it. Out of sheer frustration, in an essay on Descartes I suggested that the notion of human consciousness as 'software' that could be 'implemented' equally well on an organic brain or an optical crystal was in fact a throwback to Cartesian dualism: for 'software' read 'soul.' My tutor superimposed a neat, diagonal, luminous red line over each paragraph that dealt with this idea, and wrote in the margin (in vertical, bold-face, 20-point Times, with a contemptuous 2 Hertz flash): IRRELEVANT!

I quit philosophy and enrolled in a unit of optical crystal engineering for non-specialists. I learnt a lot of solid-state quantum mechanics. I learnt a lot of fascinating mathematics. I learnt that a neural net is a device used only for solving problems that are far too hard to be *understood*. A sufficiently flexible neural net can be configured by feedback to mimic almost any system—to produce the same patterns of output from the same patterns of input—but achieving this sheds no light whatsoever on the nature of the system being emulated.

'Understanding,' the lecturer told us, 'is an overrated concept. Nobody really *understands* how a fertilized egg turns into a human. What should we do? Stop having children until ontogenesis can be described by a set of differential equations?'

I had to concede that she had a point there.

It was clear to me by then that nobody had the answers I craved—and I was hardly likely to come up with them myself; my intellectual skills were, at best, mediocre. It came down to a simple choice: I could waste time fretting about the mysteries of consciousness, or, like everybody else, I could stop worrying and get on with my life.

When I married Daphne, at twenty-three, Eva was a distant memory, and so was any thought of the communion of souls. Daphne was thirty-one, an executive in the merchant bank that had hired me during my Ph.D., and everyone agreed that the marriage would benefit my career. What she got out of it, I was never quite sure. Maybe she actually liked me. We had an agreeable sex life, and we comforted each other when we were down, the way any kind-hearted person would comfort an animal in distress.

Daphne hadn't switched. She put it off, month after month, inventing ever more ludicrous excuses, and I teased her as if I'd never had reservations of my own.

'I'm afraid,' she confessed one night. 'What if I die when it happens—what if all that's left is a robot, a puppet, a *thing*? I don't want to *die*.'

Talk like that made me squirm, but I hid my feelings. 'Suppose you had a stroke,' I said glibly, 'which destroyed a small part of your brain. Suppose the doctors implanted a machine to take over the functions which that damaged region had performed. Would you still be 'yourself'?'

'Of course.'

'Then if they did it twice, or ten times, or a thousand times —'

'That doesn't necessarily follow.'

'Oh? At what magic percentage, then, would you stop being 'you'?'

She glared at me. 'All the old clichéed arguments —'

'Fault them, then, if they're so old and clichéed.'

She started to cry. 'I don't have to. Fuck you! I'm scared to death, and you don't give a shit!'

I took her in my arms. 'Sssh. I'm sorry. But everyone does it sooner or later. You mustn't be afraid. I'm here. I love you.' The words might

have been a recording, triggered automatically by the sight of her tears.

'Will you do it? With me?'

I went cold. 'What?'

'Have the operation, on the same day? Switch when I switch?'

Lots of couples did that. Like my parents. Sometimes, no doubt, it was a matter of love, commitment, sharing. Other times, I'm sure, it was more a matter of neither partner wishing to be an unswitched person living with a jewel-head.

I was silent for a while, then I said, 'Sure.'

In the months that followed, all of Daphne's fears—which I'd mocked as 'childish' and 'superstitious'—rapidly began to make perfect sense, and my own 'rational' arguments came to sound abstract and hollow. I backed out at the last minute; I refused the anaesthetic, and fled the hospital.

It won't be easy. He met a woman on the beach, the day I came into being. Her name is Cathy. They've slept together three times, and he thinks he loves her. Or at least, he's said it to her face, he's whispered it to her while she slept, he's written it, true or false, into his diary.

I feel nothing for her. She's a nice enough person, I'm sure, but I hardly know her. Preoccupied with my plight, I've paid scant attention to her conversation, and the act of sex was, for me, little more than a distasteful piece of involuntary voyeurism. Since I realised what was at stake, I've *tried* to succumb to the same emotions as my alter ego, but how can I love her when communication between us is impossible, when she doesn't even know *I* exist?

If she rules his thoughts night and day, but is nothing but a dangerous obstacle to me, how can I hope to achieve the flawless imitation that will enable me to escape death?

He's sleeping now, so I must sleep. I listen to his heartbeat, his slow breathing, and try to achieve a tranquillity consonant with these rhythms. For a moment, I am discouraged. Even my *dreams* will be different; our divergence is ineradicable, my goal is laughable, ludicrous, pathetic. Every nerve impulse, for a week? My fear of detection and my attempts to conceal it will, unavoidably, distort my responses; this knot of lies and panic will be impossible to hide.

Yet as I drift towards sleep, I find myself believing that I *will* succeed. I *must*. I dream for a while—a confusion of images, both strange and mundane, ending with a grain of salt passing through the eye of a needle—then I tumble, without fear, into dreamless oblivion.

I stare up at the white ceiling, giddy and confused, trying to rid myself of the nagging conviction that there's something I *must* not think about.

Then I clench my fist gingerly, rejoice at this miracle, and remember.

Up until the last minute, I thought he was going to back out again—but he didn't. Cathy talked him through his fears. Cathy, after all, has switched, and he loves her more than he's ever loved anyone before.

So, our roles are reversed now. This body is *his* strait-jacket, now . . .

I am drenched in sweat. *This is hopeless, impossible.* I can't read his mind, I can't guess what he's trying to do. Should I move, lie still, call out, keep silent? Even if the computer monitoring us is programmed to ignore a few trivial discrepancies, as soon as he notices that his body won't carry out his will, he'll panic just as I did, and I'll have no chance at all of making the right guesses. Would *he* be sweating, now? Would *his* breathing be constricted, like this? *No.* I've been awake for just thirty seconds, and already I have betrayed myself. An optical-fibre cable trails from under my right ear to a panel on the wall. Somewhere, alarm bells must be sounding.

If I made a run for it, what would they do? Use force? I'm a citizen, aren't I? Jewel-heads have had full legal rights for decades; the surgeons and engineers can't do anything to me without my consent. I try to recall the clauses on the waiver he signed, but he hardly gave it a second glance. I tug at the cable that holds me prisoner, but it's firmly anchored, at both ends.

When the door swings open, for a moment I think I'm going to fall to pieces, but from somewhere I find the strength to compose myself. It's my neurologist, Dr Prem. He smiles and says, 'How are you feeling? Not too bad?'

I nod dumbly.

'The biggest shock, for most people, is that they don't feel different at all! For a while you'll think, 'It can't be this simple! It can't be this easy! It can't be this *normal*!' But you'll soon come to accept that *it is*. And life will go on, unchanged.' He beams, taps my shoulder paternally, then turns and departs.

Hours pass. *What are they waiting for?* The evidence must be conclusive by now. Perhaps there are procedures to go through, legal and technical experts to be consulted, ethics committees to be assembled to deliberate on my fate. I'm soaked in perspiration, trembling uncontrollably. I grab the cable several times and

yank with all my strength, but it seems fixed in concrete at one end, and bolted to my skull at the other.

An orderly brings me a meal. 'Cheer up,' he says. 'Visiting time soon.'

Afterwards, he brings me a bedpan, but I'm too nervous even to piss.

Cathy frowns when she sees me. 'What's wrong?'

I shrug and smile, shivering, wondering why I'm even trying to go through with the charade. 'Nothing. I just . . . feel a bit sick, that's all.'

She takes my hand, then bends and kisses me on the lips. In spite of everything, I find myself instantly aroused. Still leaning over me, she smiles and says, 'It's over now, okay? There's nothing left to be afraid of. You're a little shook up, but you know in your heart you're still who you've always been. And I love you.'

I nod. We make small talk. She leaves. I whisper to myself, hysterically, 'I'm still who I've always been. I'm still who I've always been.'

Yesterday, they scraped my skull clean, and inserted my new, non-sentient, space-filling mock-brain.

I feel calmer now than I have for a long time, and I think at last I've pieced together an explanation for my survival.

Why do they deactivate the teacher, for the week between the switch and the destruction of the brain? Well, they can hardly keep it running while the brain is being trashed—but why an entire week? To reassure people that the jewel, unsupervised, can still stay in synch; to persuade them that the life the jewel is going to live will be exactly the life that the organic brain 'would have lived'—whatever that could mean.

Why, then, only for a week? Why not a month, or a year? Because the jewel *cannot* stay in synch for that long—not because of any flaw, but for precisely the reason that makes it worth using in the first place. The jewel is immortal. The brain is decaying. The jewel's imitation of the brain leaves out—deliberately—the fact that *real* neurons *die*. Without the teacher working to contrive, in effect, an identical deterioration of the jewel, small discrepancies must eventually arise. A fraction of a second's difference in responding to a stimulus is enough to arouse suspicion, and—as I know too well—from that moment on, the process of divergence is irreversible.

No doubt, a team of pioneering neurologists sat huddled around a computer screen, fifty years ago, and contemplated a graph of the probability of this radical divergence, versus time. How would they have chosen *one week*? What probability would have been acceptable? A tenth of a percent? A hundredth? A thousandth? However safe they decided to be, it's hard to imagine them choosing a value low enough to make the phenomenon rare on a global scale, once a quarter of a million people were being switched every day.

In any given hospital, it might happen only once a decade, or once a century, but every institution would still need to have a policy for dealing with the eventuality.

What would their choices be?

They could honour their contractual obligations and turn the teacher on again, erasing their satisfied customer, and giving the traumatised organic brain the chance to rant about its ordeal to the media and the legal profession.

Or, they could quietly erase the computer records of the discrepancy, and calmly remove the only witness.

So, this is it. Eternity.

I'll need transplants in fifty or sixty years' time, and eventually a whole new body, but that prospect shouldn't worry me—*I* can't die on the operating table. In a thousand years or so, I'll need extra hardware tacked on to cope with my memory storage requirements, but I'm sure the process will be uneventful. On a time scale of millions of years, the structure of the jewel is subject to cosmic-ray damage, but error-free transcription to a fresh crystal at regular intervals will circumvent that problem.

In theory, at least, I'm now guaranteed either a seat at the Big Crunch, or participation in the heat death of the universe.

I ditched Cathy, of course. I might have learnt to like her, but she made me nervous, and I was thoroughly sick of feeling that I had to play a role.

As for the man who claimed that he loved her—the man who spent the last week of his life helpless, terrified, suffocated by the knowledge of his impending death—I can't yet decide how I feel. I ought to be able to empathise—considering that I once expected to suffer the very same fate myself—yet somehow he simply isn't *real* to me. I know my brain was modelled on his—giving him a kind of causal primacy—but in spite of that, I think of him now as a pale, insubstantial shadow.

After all, I have no way of knowing if his sense of himself, his deepest inner life, his experience of *being*, was in any way comparable to my own.

Feminism in Philosophy of Mind

The Question of Personal Identity

Susan James

1. Introduction

A great deal of recent feminist work on philosophy of mind has been grounded on a central claim: that the key oppositions between body and mind, and between emotion and reason, are gendered. While the mind and its capacity to reason are associated with masculinity, the body, together with our emotional sensibilities, are associated with the feminine. Evidence for this view comes from at least two sources. First, overtly sexist philosophers have in the past claimed that women are by nature less capable reasoners than men and are more prone to ground their judgements on their emotional responses. These authors have been repeatedly opposed by defenders of women, whether male or female. Secondly, feminists have explored ways in which gendered oppositions are at work even in the writings of philosophers who do not explicitly differentiate the mental capacities of men and women or connect women with the bodily work of reproduction and domestic labour. By studying the metaphorical structures of philosophical texts, looking at what may appear to be digressions from the main line of argument, and paying attention to examples, they have identified persistent patterns of association running through the history of philosophy. These patterns can fluctuate from century to century, from author to author, from work to work, and even from paragraph to paragraph, but they keep cropping up. They indicate that the terms associated with the feminine are persistently marginalized by comparison with those associated with masculinity, as when the rational powers of human beings are habitually regarded as more valuable than their emotional skills.[1]

In the light of this analysis, many feminists have worked to develop philosophical positions which do not devalue the symbolically feminine. They have done so by unsettling the hierarchical relations between mind and body, and between reason and emotion, approaching their task in various overlapping ways. Sometimes they have criticized existing, influential theories of body and mind; sometimes they have reconceptualized particular topics within the philosophy of mind; and sometimes they have drawn on the work of authors who have written 'against the grain.'

A prominent example of the first approach has been the engagement of feminist philosophers with the phenomenological tradition, and particularly with the work of Merleau-Ponty.[2] However, by far the most striking case of this type of constructive criticism is to be found in the troubled relationship between feminism and psychoanalysis. In the anglophone world, this originated in a sequence of critical readings of Freud,[3] and subsequently developed into a debate which both takes issue with the psychoanalytic tradition, and deploys its resources. Diverse contributors to the discussion have drawn not only on the ideas of Freud himself, but also on those of Klein, Winnicott and Lacan to explain aspects of sexual difference and to reconsider the oppositions mentioned above.[4] Interest in Lacan, and indeed in other strands of psychoanalytic thought, has been stimulated by the work of some extremely influential French authors, notably Luce Irigaray and Julia Kristeva.[5]

Turning to the second approach, feminist writers have directly addressed the opposition between body and mind, in an effort to reveal how the body is tacitly marginalized in philosophy and to find ways of reinstating it. Much of this work aims to question the distinction between the mental and the physical by showing how mind and body interrelate, and how the body contributes to, and is implicated in, thought.[6] A number of influential contributors to this project have focused on the distinction between sex and gender. Originally coined to mark a division between the biological and social characteristics differentiating men and women, this distinction has been repeatedly questioned, to the point where there is now widespread doubt as to whether it is fruitful to try to keep these two groups of properties apart. Querying the idea of the purely bodily casts doubt on the existence of a clear division between the mental and the physical, while emphasizing the social challenges the sufficiency of an opposition between body and mind.[7]

From Miranda Fricker & Jennifer Hornsby, eds., *The Cambridge Companion to Feminism in Philosophy* (Cambridge University Press, 2000). Reprinted by permission of the publisher.

Directing their attention to the relation between reason and emotion, feminist philosophers have argued that emotion is integral to reasoning, and have brought out some of the ways in which emotion traverses the divide between mind and body.[8] In addition, they have taken a step which characterizes a good deal of feminist work in philosophy of mind, and is one of its claims to originality. By charting the ways in which particular emotions are held to be appropriate in men and inappropriate in women, or appropriate in women and inappropriate in men, they have linked together issues which have generally been held apart, and shown how political philosophy and philosophy of mind are connected.[9]

2. Personal identity

Several of these themes can be traced in contemporary feminist writing about personal identity, which has tended to draw on the insights of psychoanalysis and postmodernism to explore the ways in which selves are embodied, discontinuous, malleable and socially constructed.[10] At the same time, anglophone theorists of personal identity have continued to develop a conception of the self which revolves around a distinction between the psychological and the bodily, and a related notion of psychological continuity.[11] It is tempting to suppose that these two groups are addressing different questions: that feminists are for the most part interested in the variety of ways in which identity can be moulded, lived, or transformed; and that theorists of personal identity are concerned with the prior question of what it is to have an identity at all. But this suggested division of labour is too simple. Feminist explorations of the self are, among other things, attempts to depart from the symbolically masculine character of much of philosophy, and their concern with embodiment, discontinuity and social construction is driven by a desire to avoid reiterating the hierarchical oppositions outlined in the preceding section. By embodying the self, they aim to undo the deeply rooted association between the self and the masculine mind; by emphasizing discontinuity, they aim to put pressure on the cultural alliance between unity and masculinity. From a feminist perspective, therefore, the continued dependence of personal identity theorists on various oppositions that feminist philosophy aims to dismantle is at least suspicious. In this chapter I shall explore some of the grounds for this suspicion, and suggest ways in which it is well-founded.

Within the analytic tradition, discussion of persons focuses largely on the question: what criteria have to be satisfied in order for it to be true that a person at t1 survives at t2? Or: what criteria have to be met for a person at t1 to be the same person at t2?[12] Until recently, these were generally taken to be questions about personal identity, and it was widely assumed that any relation specifying continuing personhood would have to share some key features of the identity relation, such as transitivity and being one-one. Feminists who have argued that philosophy places too much emphasis on identity, and uses it to maintain the system of binary oppositions which exclude the feminine, might have found this objectionable. But in any case, Derek Parfit's work has prompted a reconsideration of this claim. What matters, he has suggested, is survival.[13] And if persons can survive without being identical, the way is open to allow that survival may be a matter of degree. We reach the possibility of a more flexible conception of selfhood which is consonant with at least some feminist arguments.

At the same time, contributors to the debate have found it helpful to distinguish two criteria for continuing personal identity—bodily continuity and psychological continuity—and in this way to separate body and mind. Among feminists, this sort of approach is widely regarded as worthy of scrutiny, as it is sometimes the prelude to an attempt to marginalize the body, and with it the symbolically feminine. In this particular case it is undoubtedly the prelude to a manoeuvre which reinforces the mind/body divide, namely the construction of thought experiments which press these two apart. In the last few years, a good deal of weight has been placed on imaginary examples which suggest that psychological as opposed to bodily continuity is what constitutes a person's survival. One kind of example, in particular, has been crucial in securing this view: the much-cited cases in which, by some means or other, one person's character and memories are transplanted into a second person's body.[14] Although other scenarios such as fission and fusion are also appealed to,[15] transplant cases are a crucial resource on which theorists of various persuasions rely, and are used to create a framework within which different accounts of survival can be discussed.

To make a case for the view that the debate about personal identity marginalizes the

feminine, and is one of the ways in which philosophy privileges the symbolically masculine over its feminine counterpart, I shall concentrate on these examples. I shall not discuss the relative merits of psychological and bodily continuity as conditions of survival, nor shall I consider the relation between survival and identity. Instead, I shall try to show how imaginary examples of character transplant are used to sustain a symbolically masculine conception of personhood. I shall take up four points: one about the delineation of character; a narrower one about memory; a third about the role of the social world in sustaining identity; and a fourth about identity and male sexual power.

3. Delineation of character

Imaginary cases in which one person's character is transplanted into another person's body generally assume that character has to be lodged in a material body of some sort. It may be a whole human body, a brain, or half a brain. The body in question may be inorganic, as when an imaginary machine stores the information from one brain and prints it off in another.[16] But in all these versions the body is thought of as a container or receptacle for character. The brain figures as a container in which a person's psychological states can be preserved, and the body figures as a more elaborate receptacle for the brain. Equally, a machine which copies the information from one brain and prints it into another is a receptacle for storing psychological states.

Several contributors to the literature on personal identity acknowledge that thinking of the body as a receptacle may be an excessive oversimplification, but brush this thought aside. In "The Self and the Future," for example, Bernard Williams notes that body swapping between people of different sexes may be hard to imagine, but comments 'Let us forget this,'[17] so turning his back on a point he makes elsewhere, that it may be impossible for an emperor to express his personality when his body is that of a peasant.[18] Other writers, such as Noonan, note the problem, but bypass it by specifying that the bodies in question are either only numerically distinct, or extremely similar.[19] Any characteristics that might enable the body to disrupt the psychological continuity of the character transplanted into it are removed, with the result that bodies are regarded, for the purposes of the experiment, as uniform. They do of course differ in various ways, but these differences are held to be irrelevant.

Making the body anonymous in this way simultaneously affirms a particular view of what character is. The things that really matter about a person's character, the traits which constitute their psychological continuity, do not depend on their having a particular body, or a body with any particular properties. Anthony Quinton makes this point explicitly. 'As things are,' he writes, 'characters can survive large and even emotionally disastrous alterations to the physical type of a person's body . . . Courage, for example, can perfectly well persist even though the bodily conditions for its obvious manifestation do not.'[20] Courage, perhaps, but what about dexterity? Patience, perhaps, but what about delight in one's sexuality? (It would be interesting to consider whether all the traditional virtues can be construed as independent of the body in this way.) Quinton's argument exemplifies a tendency which runs through imagined cases of character transplant—a tendency to rely on a conception of character or psychological continuity which serves to emphasize, and even create, a division between the psychological and the bodily. Properties which do not fit neatly into the category of the psychological are held to be marginal or irrelevant to character. Then, if continuity of character is taken to be what matters in survival, merely bodily states become irrelevant to survival.

4. Memory

Partly because the states that contribute to psychological continuity are specified as states that are not bodily, theorists of personal identity are able to be both non-committal and inclusive about what exactly they are. Lewis, for example, regards this as a question of detail,[21] and Noonan claims that 'in general *any* causal links between past factors and present psychological traits can be subsumed under the notion of psychological connectedness.'[22] However, a central role is often given to memories as states which give us access to our pasts, and secure our sense of temporal continuity. How must memory be conceived if it is to fulfil this function, while leaving intact the division between body and character?

At least, I suggest, as a storehouse of recollections able to survive bodily vicissitudes. Take the case of Adam. Whatever happens to him—even if he has one of his ribs removed,

even if his body changes beyond recognition, even if God refuses to recognize him—he will still be able to think of a sequence of things he did and things that happened to him as *his* actions and experiences. More particularly, changes in his body will not interfere with this capacity. For example, even when he is weak and wasted he remembers that he took the apple from Eve as a strong young man. Why, then, should this capacity not endure in the imaginary case where Adam's character is transplanted into a different body?

There are some obvious exceptions to the view that memory is unaffected by bodily vicissitudes. For instance, brain damage may make Adam amnesiac, and if his character is transplanted into a body with a damaged brain, it is not obvious that his memories will survive. A more interesting example is provided by cases of physical violation such as rape, other forms of torture, or malicious attack, which often have a profound impact on memory. In an illuminating paper, Susan Brison makes the point that experiences like these do not simply add to the victim's stock of memories, as a camera operator might shoot another few feet of film, nor are they safely lodged in the mind, as the camera operator might store the exposed film in a tin.[23] First, memories of trauma are in many cases closely tied to the body, indeed are *in* the body, and manifest themselves in physical states as much as in psychic ones. Here any neat separation between bodily states and memory as the bearer of psychological continuity seems to break down. To press a tasteless question, would a trauma victim retain her memories if her character were transplanted into a different body? Secondly, trauma destroys or alters existing memories, so that people who have been subjected to extended torture or deprivation lose conscious memories of their own pasts, and lose, too, the easy sense of continuity that memory is here supposed to provide. Their time scale may shrink so that their memories of their own experiences become mainly short-term ones. And the continuity of memory may be punctured and jumbled by uncontrollable, nightmarish recollections.

Writers on personal identity usually try to take account of the loops, breaks and fade-outs in our memories by emphasizing that psychological continuity does not require a single sequence of memories, but only a sequence of overlapping sequences. Furthermore, it is not constituted by memory alone. Where memory breaks down, other continuities such as those in a person's desires, intentions, or hopes can take over. The fact that trauma victims lose memories therefore need not imply that they lose psychological continuity. However, Brison's discussion identifies one of the limitations of such an approach. This way of thinking about continuity suggests that, when memory fails, other psychological states remain unchanged and serve as the guarantors of personhood. But trauma victims do not just lose their memories of past events or actions. They lose the pattern of memory in which their expectations, emotions, skills, desires, and so on are rooted, so that loss of memory is, in these cases, part of a broader destruction of character. The ability to enjoy dancing, for instance, is grounded on remembered physical skills (how to tango), expectations derived from past experience (that one will be safe), emotional dispositions (taking pleasure in music), the confidence that one can keep one's own memory under control, and so on. When all these are gone, enjoying dancing will be gone also. And so for other character traits.

If, as much discussion of personal identity assumes, memory is to be one of the guarantors of psychological continuity, and if psychological continuity is to be separable from bodily continuity, memory must be interpreted in a particular and selective way. Memories in the body have to be set aside in favour of those which appear to have no bodily aspects; and it has to be assumed that the impact of memory loss on other character traits is sufficiently limited for psychological continuity to survive. It is arguable that these are not very contentious assumptions. But they nevertheless help us to see that the division between body and character, around which imaginary transplant cases are organized, can only be sustained if the traits constituting character are laundered, and all traces of the body washed away. The purified conception of 'the psychological' which emerges then appears as an unsullied self for which the body is simply a convenient receptacle.

5. Social circumstances

The two steps we have examined—the expelling of everything bodily from the mind, and the simultaneous devaluation of the bodily—are familiar to feminists, many of whom have read them as an attempt to demarcate the masculine from the feminine and exclude the latter from philosophy. We can find further traces of

this way of proceeding in discussions of personal identity if we focus on another curious feature of the persons around whom debate rages—their complete lack of any history or social context. As we have seen, the key question that concerns philosophers is what it takes for x at t1 to survive at t2. This assumes that we start out with a fully-fledged person, which is why I've called him Adam. And it assumes that in ordinary circumstances (if he doesn't die) he will survive until t2. Philosophers who regard psychological continuity as what matters in survival thus assume that psychological continuity is a property of normal human beings.

To take it for granted that Adam at his creation is a person is to suppose that at that point he has both a body and a character—a suitably integrated set of memories, emotions, desires and so on. The expectation that in normal circumstances he will survive to be expelled from Paradise has built into it the expectation that he possesses the means to maintain his character in some body or other, to satisfy the demands of psychological continuity. These are large assumptions which exclude a good deal. The first excludes the fact that character, in the sense of the ability to understand oneself as the subject of diverse psychological states, is not a birthright, but the fruit of a child's relations with the people who care for him or her. Theorists of personal identity appear to take a Lockean view of the genesis of character: once Adam is created, or once a baby reaches a certain stage, memory starts to roll and an integrated character develops. In doing so they exclude from consideration some of the ways in which the self is dependent on others, particularly on its mother figure. At the same time they make it unnecessary to consider whether features of the process by which the self is constituted may effect its subsequent continuity. The second assumption has complementary consequences: it brackets the question of whether the maintenance of psychological continuity also depends on social relations.

From Freud onwards, writers in the psychoanalytic tradition have elaborated the view that a child's experiences are not initially integrated or continuous, and are not initially the experiences of an individuated self. Coming to understand itself as separate from its mother figure, and becoming able to claim its experiences as its own, is for a child a process in the course of which it becomes able to locate its experiences in its own body. As a number of feminists have stressed, both Freud and Lacan describe the ego as a psychical mapping of the libidinal intensities of the body, a mental projection not of the actual body, but of the body as a kind of emotional map.[24] Freud's ideas are elaborated in Lacan's argument that, during the mirror stage, the child forms an image of its own body as it is represented for it by the images of others, and by its own reflection in a mirror. This image, which is of the body as a whole, forms a sort of provisional identity. It is itself a precondition of the more stable symbolic identity the child acquires as the result of the resolution of the Oedipus complex. And it survives the Oedipus complex as the ego ideal, a model of bodily integrity. Work on body images suggests that they make an important contribution to psychological continuity. During the mirror stage the child embarks on the process of coming to understand itself as situated in the space occupied by its body; or, to put the point differently, embarks on the process of acquiring a stable emotional investment in its body. Only once it has a body image can it understand its body as 'mine,' and only then can it possess a perspective on the world.[25]

The self for whom psychological continuity is a possibility therefore has to be created through a series of interactions between the child, people around it, and the broader culture in which it lives.[26] Equally, psychological continuity has to be sustained, and social circumstances can either foster or damage it. To return to Susan Brison's argument, trauma victims who describe the selves they were as dead, or beyond recognition, provide searing evidence of the ways that continuity can be shattered. As well as losing the memories and character traits which defined them, they may have lost the ability to inhabit fully the lives they are now living. Brison quotes a poem by Charlotte Delbo about her return from Auschwitz to Paris:

> life was returned to me
> and here I am in front of life
> as though facing a dress
> I cannot wear.[27]

To recover the sense of subjectivity that personal identity theorists so often take for granted, such people need to recover the ability to care about themselves and the world, to feel emotions that, as Brison puts it, are more than counterfactual.[28] Others can play a crucial part in this process. By listening to, and recognizing, the victims of trauma, others seem to be able to help them piece together their memories into narratives with which they can

identify, and master the troubling bodily manifestations of memory which further disrupt the self. Extreme cases like these suggest that psychological continuity has a social dimension insofar as it depends on recognition by others. When recognition is withdrawn, the emotional investment in our memories and characters that holds the self together may be weakened, so that, to varying degrees, we suffer a kind of depersonalization—an inability to feel that our experiences are our own, and a subsequent inability to integrate and order them.

The view that psychological continuity has to be created and sustained has some impact on the personal identity theorists' assumption that bodily and psychological continuity are conceptually separable. The arguments I have just sketched help us to elaborate an account of what is left out in the imaginary cases where it is assumed that psychological continuity would survive body transplant. Suppose we assume that psychological continuity does depend on the possession of a body image, and on an emotional investment in it. Is it now so obvious that the features of the body into which a character is transplanted are irrelevant to its survival? To dramatize the issue in a manner typical of this philosophical literature, what about a female fashion model whose character is transplanted into the body of a male garage mechanic? Might she not find it impossible to reconcile her body image with the body that had become hers, and suffer such a level of dislocation that she became unable to locate her experiences in that body? At the limit, might she not experience the depersonalization suffered by some psychotics, who lose interest in the whole body and do not invest any narcissistic libido in the body image? Their self-observations seem viewed from the perspective of the outsider and they display no interest in their own bodies.[29] Suppose, by contrast, we imagine a character whose body is transplanted into that of her identical twin. The point is that she remains psychologically continuous (if she does) because the body that is now hers has properties which make it possible for her to live in it as her own. Psychological continuity is not independent of the body. It is a feature of embodied selves.

If recognition makes a difference, the degree of a person's psychological continuity may also depend on social circumstances. To return to the case of the model, will her friends and lovers continue to recognize and affirm her? Will she be able to find anyone able to believe her story and hear her out? Anthony Quinton touches optimistically on the first point. 'In our general relations with other human beings their bodies are for the most part intrinsically unimportant. We use them as convenient recognition devices enabling us to locate the persisting characters and memory complexes . . . which we love or like. It would be upsetting if a complex with which we were emotionally involved came to have a monstrous or repulsive physical appearance . . . But that our concern and affection would follow the character and memory complex . . . is surely clear.'[30] Quinton is aware that this may not quite settle the argument, and addresses the looming objection that some personal relations, such as those 'of a rather unmitigatedly sexual type,' might not survive a change of body. But here, too, he resolves the problem confidently. 'It can easily be shown that these objections are without substance. In the first place, even the most tired of entrepreneurs is going to take some note of the character and memories of the companion of his later nights at work. He will want her to be docile and quiet, perhaps, and to remember that he takes two parts of water to one of scotch, and no ice . . . As a body she is simply an instrument of a particular type . . .'[31] This solution to the problem employs the strategy we have already examined: it resolutely divides psychological properties from bodily ones and insists that the former are what matter in recognition. The wish to be loved for oneself alone and not for one's golden hair is simply granted. What this solution does not countenance, however, is the possibility that a person's ability to sustain psychological continuity may depend on other people recognizing and affirming the properties and potentialities of their embodied selves, and that where this possibility is removed, their psychological continuity may be damaged.

6. Marginalizing the symbolically feminine

We can now see more clearly that when personal identity theorists specify that characters are transplanted into bodies identical with the ones they had before, they are not introducing innocent simplifications. Instead, they are covering up and discounting ways in which psychological continuity is woven into the histories of our embodied selves. However, this is not the end of the matter. A theorist of personal identity may concede that psychological continuity has to be created, and that in extreme cases such

as psychosis it can be destroyed. But he or she may nevertheless maintain that, in all ordinary cases, once psychological continuity is created, it survives. We see this, for example, in the testimony of the victims of extreme and extended trauma. While they may not remember much about their earlier lives, and may now lack well-defined characters, they identify with their past selves and speak about them in the first person (albeit sometimes rather oddly as when they say things like 'I died there' or 'I shall always miss myself as I was then'). We see it, too, in cases of physical mutilation where, although the body image usually takes some time to adjust, people do not lose all sense of who they are.[32] Only in pathological conditions such as psychosis and multiple personality does the self really fragment. So, putting these last cases aside, are we not right to posit a sense of psychological continuity which is independent of both bodily and social vicissitudes, or to imagine that this sense of continuity could survive if a character were transplanted from one body into another?

The arguments I have offered aim to show that, once we strip this imaginary situation of features which function to make it appear unproblematic, the kind of continuity that can be relied on is comparatively attenuated. All we are able to assume is that the transplanted character is able to locate its experiences in its new body, and that it remains sufficiently integrated to claim some memories as its own. We need not assume that it has much emotional investment in its memories. Nor need we assume much continuity of other character traits. Psychological continuity features here as a slender lifeline which enables the transplanted person to say to themselves 'I know that such and such happened to me and that I am so and so' and just about to believe it.

The personal identity theorist must be prepared to argue that this minimal level of continuity is sufficient to sustain the claim that we can fruitfully explore the question of what is involved in survival by playing off bodily and psychological continuity against one another. It seems to me, however, that the attractions of psychological continuity as a separable component of survival have been considerably reduced. Let me labour this point. Before, we were imagining that, transplanted into a new body, I would feel pretty much the same as I do now, would be able to continue the projects I have now, would be no less committed to my future than I am now, would have the memories and characteristics I now possess, and would retain the relations with other people that, so it seems to me, make life worth living. Now we imagine a situation in which it is much less clear what transplant will be like, and in which it may give rise to psychic and physical pain comparable, perhaps, to the pain of torture which looms so large in one of the problem cases constructed by Williams.[33] I may lose many memories and character traits, so that my hold on my own past is tenuous and emotionally numbed, and my grasp of who I now am is fractured and confused. I may lose the affection and even recognition of the people who matter to me, and also the capacity to form new relationships. I may be unable to pursue my projects or embark on new ones, and may have very little emotional investment in the life I am living.

Some theorists of personal identity would, I suspect, insist that as long as there remains a thread of continuity between the pre- and post-transplant selves, we have a case for the conclusion that they are the same person. The barest 'I' is enough to hold the self together and to underwrite an approach to the problem that separates psychological and bodily continuity. But in the light of the sorts of difficulties I have discussed it seems reasonable to ask: Why cling to this doctrine? Why deploy such resources of imagination to prise the bodily and the psychological apart? And why go to such lengths to protect psychological continuity from the effects of the body and the rest of the world?

At this point a reader might object that these questions misrepresent the current debate. Contemporary theorists of personal identity, it might be claimed, are by no means agreed that psychological continuity is essential, or even important, to personhood, and many of their accounts emphasize the centrality of the body. This is undoubtedly true. However, the approach I have been discussing is extremely influential, and continues to shape our understanding of what the problem of personal identity consists in.[34] As long as this much is conceded, the questions I have posed remain pertinent.

Feminists who have addressed these questions have frequently drawn on a conception of the self which is set over against, though not completely irreconcilable with, the view of personal identity we have been examining, insofar as it holds that there is an important aspect of the psyche, the unconscious, which this view neglects. To accept that the unconscious is at work when we philosophize is to accept that the psychological discontinuities so evident in

pathological cases are present to some degree in all of us. Some aspects of the self are simply not picked up by accounts which emphasize psychological continuity, and the decision to discount these may itself have unconscious motivations. Taking the unconscious into account, then, feminist philosophers have explained the prominence of views which regard the body as unimportant to identity in various ways. Some have argued for the view that, in European culture, the mind is associated with masculinity and the body with femininity. One term can stand in for, or symbolize, the other. Philosophers (most of them men) have employed these associations. They have assumed (often unconsciously) that personal identity is male identity, and have developed accounts in which the symbolically masculine mind is given priority over the body.[35] Other writers have provided psychological explanations for this downgrading of the symbolically feminine. When male personal identity theorists construct imaginary examples which separate the bodily from the psychological, they resolve in fantasy the always-unresolved conflicts of the Oedipus complex—the separation of a male child from his mother figure, and his subsequent identification with his father. In establishing and maintaining a firm boundary between the maternal body and the paternal mind, they deny their own unconscious desire to be reunited with the mother figure. And in fixing on psychological continuity as the mark of identity, they construct a picture in which masculinity and selfhood coincide.[36] A further aspect of the transplant fantasy also serves to exclude the feminine. By positing fully fledged persons whose history is irrelevant to the problem at hand, male philosophers imagine for themselves a condition of self-sufficiency, from which their indebtedness to a mother figure, or indeed to anyone else, is excluded.

These two types of explanation (one cultural, the other psychological) have a good deal in common. Both rest on the claim that philosophers (male and female) are themselves psychologically discontinuous, in the commonplace sense that their unconscious fears and desires play a part in determining the way they formulate and argue about problems, and the sorts of arguments they find persuasive, although this is not an aspect of their philosophizing over which they have conscious control. Moreover, both assume that particular associations at work in our culture continue to play a significant part in shaping our philosophical beliefs. According to the first kind of view, symbolic associations help to explain the fact that we privilege some terms over others. According to the second, these symbolic associations are themselves embedded in the psychological processes that form sexual identity.

Over the last two decades, feminist philosophers have amassed a range of evidence for both the explanatory hypotheses I have sketched. However, it remains to ask what internal support we can find for the view that theorists who equate personal identity with psychological continuity are upholding (however unconsciously) a masculine conception of identity. I have assumed, uncontentiously I hope, that we sometimes find clues to the unconscious in questions that hover round the margins of a text, so that when Williams or Noonan allow that transplant from one body into a very different one might be difficult, and then immediately put the problem aside, it is probably worth looking further.[37] I have also assumed—and Williams and Quinton make this explicit—that what they are putting aside here is the issue of sexual identity.[38] To return to the fantasy of character transplant, there are in principle a variety of ways of thinking about the case of a male character transplanted into a female body. Maybe it would be the ideal sex-change operation. Maybe it would condemn the resulting person to the unhappy condition of someone who desperately wants a sex change operation. Maybe it would produce psychological breakdown. As we have seen, most writers block off exploration of lines of thought like these, which require us to think of the people concerned as embodied, in their investigations of personal identity. Why? Perhaps because they take it that the identity of a person is the identity of a male. Perhaps because an unconscious fear of jeopardizing their sexual identity prevents them from doing so, and helps to direct them towards an approach which brackets the body and concentrates on the mind.

It may be helpful to consider what kinds of criticism I have offered of the view that personal identity consists in psychological continuity. In the preceding sections of this essay I have voiced some objections to this analysis which *can* be assessed independently of any claims about gender as arguments to the effect that authors who appeal to a particular kind of thought experiment rely on an inadequate conception of the self. The limitations of the conception they employ undermine not only the particular conclusions they draw from their

thought experiments, but their very approach, which works with an oversimplified conception of memory, neglects the social construction of the self, and is insensitive to the ways in which selves are embodied. At the same time, however, I have claimed that the issue of gender is woven into arguments which rely on fantasies of brain transplant, and to bring this out I have asked what is going on when philosophers advance them. What is being said, explicitly and implicitly, and why? One of the things going on, so I have suggested, is that a symbolically masculine account of identity is being unself-consciously articulated. A sceptically inclined reader may still wish to ask whether this diagnosis amounts to a criticism beyond those set out in the first part of the chapter. What is wrong with the symbolically masculine account, other than the fact that it suffers from the deficiencies just summarized?

To answer this question, it is helpful to distinguish the type of criticism which pinpoints a particular flaw in a position from the type which indicates the shortcomings of an approach. The diagnosis I have offered is of the latter kind. Its critical force rests on the assumptions that we are in search of philosophical interpretations that answer to our experience and acknowledge the complexity of our lives, and that, in the case of personal identity, part of this complexity lies in sexual identity. Theories which neglect or disavow sexual difference therefore cut themselves off from an important set of issues, and in doing so render themselves philosophically impoverished. To show how this occurs is not, of course, to specify what a feminist analysis of personal identity would be like, or to explore how a focus on sexual difference alters our understanding of the relation between personhood and embodiment, though many of the works cited throughout the chapter undertake these very tasks. My aim has been to articulate some of the features of an analytical approach to personal identity which leave feminist philosophers dissatisfied, and which explain the fact that their work has developed in different directions.

7. Identity and social power

The symbolic gendering of the opposition between body and mind, on which I have so far concentrated, has provided an exceptionally fruitful focus for feminist research. Nevertheless, it is important not to assume too readily that the body always figures as feminine and the mind as masculine,[39] or to take it for granted that gender is exclusively associated with these terms. Some theorists, I have been arguing, locate personal identity in a mind which they interpret as masculine; but there is also evidence that a man's continuing identity is sometimes implicitly understood to depend on his ability to control a woman. Here the issue is not how the 'components' of a person are gendered, but how the relations between people of different sexes bear on the problem of identity. If social relations can secure or destroy continuing identity, as I suggested earlier on, they will provide another area in which identity and gender intertwine.

This motif is central to some works of literature. For example, in Janet Lewis's novella, *The Wife of Martin Guerre*,[40] Martin Guerre leaves his village and family and does not come back. Eight years later he returns—or rather, an impostor arrives, who slips into Guerre's place and takes up the life he had left behind. Some time goes by before Guerre's wife, tortured by the belief that the impostor is not her husband, and that she is an adulteress, confesses her suspicions, and the impostor is brought to trial. Just as judgement is about to be announced, the original Martin Guerre walks into the court room, and the impostor is punished with death. In this narrative, it becomes important to establish the impostor's identity because he is usurping Guerre's sexual rights over his wife, or to put it another way, because Guerre has lost control over her. She is out of his control, and her independence of him is part of what threatens to obliterate Guerre's social identity, insofar as it is one of the conditions that allow the impostor to 'become' him. The trial restores both Guerre's identity and sexual order.

We find the same link between identity and male sexual power in Balzac's story about Colonel Chabert[41] who, when the tale begins, has been listed among the casualties of Napoleon's Russian Campaign. His name has appeared on the list of valiant heroes who sacrificed their lives for France, his wife has remarried, and his house has been sold. But in fact the Colonel has survived, and after several years returns to Paris, determined to reclaim his wife. Once again, loss of identity is linked to loss of control over a woman, and his desire to have his wife back is what drives the Colonel to explain his plight to a young lawyer, who takes up his case and tries to negotiate a settlement. In the course of the negotiations the Colonel comes to see

that his wife is a ruthless and avaricious woman who will never return to him, and has never loved him anyway, and renounces his desire to reclaim her. But the recognition that he cannot possess her destroys him, and in the final scene the lawyer comes across him, unkempt and listless, sitting on a log beside the road staring vacantly into space. Here, loss of power over a woman is associated not just with loss of social identity but with psychological discontinuity. To be sure, Colonel Chabert is deprived of his social identity; but he loses more than this, and although the man sitting on the log may know who he is, his discontinuity with his past self prevents him from functioning.

8. Conclusion

When theorists of personal identity focus on psychological continuity as the stronghold of the self, and construe psychological continuity as independent of bodily continuity, they secure only a self which would in other circumstances be regarded as pathologically disturbed. This is, to be sure, a self of sorts, and one consonant with the problem 'What is it to survive?' which already carries connotations of minimal continuity, of enduring against the odds and in the face of obstacles. Perhaps the question we should be addressing, then, is why the analytical philosophical tradition has been so concerned to explore and defend this minimal notion of survival, and hence personhood. Part of the explanation, I have suggested, lies in cultural constructions of masculinity and femininity which are at work in the unconscious, and consequently in philosophy. At the heart of identity lies the issue of sexual identity, and with it the desire of a male-dominated tradition to secure the masculinity of the subject and the subordination of women. This commonplace drama is played out in various philosophical arenas, but is worked through with particular intensity in the problem of identity itself.[42]

NOTES

1. Carolyn Merchant, *The Death of Nature: Women, Ecology and the Scientific Revolution* (San Francisco: Harper & Row; London: Wildwood House, 1980); Elizabeth Spelman, "Woman as Body: Ancient and Contemporary Views," *Feminist Studies* 8, no. 1 (1982): pp. 109–31; Genevieve Lloyd, *The Man of Reason: 'Male' and 'Female' in Western Philosophy* (London: Methuen, 1984), and 'Maleness, Metaphor and the "Crisis" of Reason,' in *A Mind of One's Own. Feminist Essays on Reason and Objectivity*, Louise M. Antony and Charlotte Witt, eds., (Boulder, CO: Westview Press, 1993), pp. 69–83; Evelyn Fox Keller, *Reflections on Gender and Science* (New Haven: Yale University Press, 1985); Luce Irigaray, *An Ethics of Sexual Difference*, Carolyn Burke and Gillian C. Gill, trans., (Ithaca, NY: Cornell University Press, 1985); Michèle Le Dœuff, *The Philosophical Imaginary* (London: Athlone, 1980); Tina Chanter, *Ethics of Eros: Irigaray's Rewriting of the Philosophers* (London: Routledge, 1995); Penelope Deutscher, *Yielding Gender: Feminism, Deconstruction and the History of Philosophy* (London: Routledge, 1997).
2. Sharon Sullivan, "Domination and Dialogue in Merleau-Ponty's *Phenomenology of Perception*," *Hypatia* 12, no. 1 (1997): pp. 1–19; Judith Butler, 'Sexual Ideology and Phenomenological Description: A Feminist Critique of Merleau-Ponty's Phenomenology of Perception,' in *The Thinking Muse: Feminism and Modern French Philosophy*, Jeffner Allen and Iris Marion Young, eds., (Bloomington: Indiana University Press, 1989), pp. 85–100; Iris Marion Young, "Throwing Like a Girl: A Phenomenology of Feminine Body Comportment, Motility, and Spatiality," in *Throwing Like a Girl and Other Essays in Feminist Philosophy and Social Theory* (Bloomington: Indiana University Press, 1990), pp. 141–59.
3. Juliet Mitchell, *Psychoanalysis and Feminism* (Harmondsworth: Penguin, 1974); Jane Gallop, *Feminism and Psychoanalysis* (London: Macmillan, 1982).
4. Nancy Chodorow, *The Reproduction of Mothering* (Berkeley: University of California Press, 1978); Dorothy Dinnerstein, *The Mermaid and the Minotaur* (New York: Harper and Row), published in the UK under the title *The Rocking of the Cradle and the Ruling of the World* (London: Souvenir Press, 1978); Teresa Brennan, ed., *Between Feminism and Psychoanalysis* (London: Routledge, 1989) and *The Interpretation of the Flesh. Freud and Femininity* (Routledge: London, 1992); Jessica Benjamin, *The Bonds of Love: Psychoanalysis, Feminism and the Problem of Domination* (London: Virago, 1990); Elizabeth Grosz, *Jacques Lacan: A Feminist Introduction* (London: Routledge, 1990). There is a list of Further Reading in Feminism and Psychoanalysis at p. 266.
5. Luce Irigaray, *Speculum of the Other Woman*, Gillian C. Gill, trans., (Ithaca, NY: Cornell University Press, 1974); Margaret Whitford, *Luce Irigaray. Philosophy in the Feminine* (London: Routledge, 1991); Julia Kristeva, *Black Sun. Depression and Melancholia*, Leon S. Roudiez, trans., (New York: Columbia University Press, 1989); Kelly Oliver, ed., *The Portable Kristeva* (New York: Columbia University Press, 1997).
6. Rosi Braidotti, *Patterns of Dissonance* (Cambridge: Polity Press, 1991); Elizabeth Grosz, *Volatile Bodies* (Bloomington: Indiana University Press,

1994); Moira Gatens, *Imaginary Bodies* (London: Routledge, 1996); Christine Battersby, *The Phenomenal Woman* (Cambridge: Polity Press, 1998).

7. Simone de Beauvoir, *The Second Sex*, H. M. Parshley, trans. and ed., (Harmondsworth: Penguin, 1972); Toril Moi, *Simone de Beauvoir: The Making of an Intellectual Woman* (Oxford: Blackwell, 1994); Ann Oakley, *Sex, Gender and Society* (London: Temple Smith, 1972); Judith Butler, *Gender Trouble: Feminism and Subversion of Identity* (London: Routledge, 1990) and *Bodies that Matter: On the Discursive Limits of Sex* (London: Routledge, 1993); Gatens, *Imaginary Bodies*; Evelyn Fox Keller, "Gender and Science: An Update," in *Secrets of Life, Secrets of Death: Essays on Language, Gender and Science* (New Haven: Yale University Press, 1985).
8. Alison Jaggar, "Love and Knowledge: Emotion in Feminist Epistemology," in *Women, Knowledge and Reality: Exploration in Feminist Philosophy*, A. Garry and M. Pearsall, eds., (London: Routledge, 1996), 2nd ed., pp. 166–90.
9. Sandra Lee Bartky, "Shame and Gender," in *Feminity and Domination: Studies in the Phenomenology of Oppression* (London: Routledge, 1990), pp. 83–98; Naomi Scheman, "Anger and the Politics of Naming," in *Engenderings: Constructions of Knowledge, Authority and Privilege* (London and New York: Routledge, 1993), pp. 22–35; Elizabeth Spelman, "Anger and Subordination," in *Women, Knowledge and Reality*, Garry and Pearsall, eds., 1st ed., pp. 263–73; Sue Campbell, *Interpreting the Personal: Expression and the Formation of Feelings* (Ithaca, NY: Cornell University Press, 1997).
10. See, for example, Rosi Braidotti, *Nomadic Subjects* (New York: Columbia University Press, 1994); Butler, *Gender Trouble*; Gatens, *Imaginary Bodies*; Grosz, *Volatile Bodies*; Young, *Throwing like a Girl.*
11. See, for example S. Shoemaker, "Personal Identity, A Materialist View," in *Personal Identity*, S. Shoemaker and R. Swinburne (Oxford: Blackwell, 1984); Bernard Williams, "The Self and the Future," *Philosophical Review* 79 (1970): pp. 161–80, reprinted in *Problems of the Self* (Cambridge: Cambridge University Press, 1973), pp. 46–63; Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984); David Lewis, "Survival and Identity," in *Philosophical Papers*, vol. 1 (Oxford: Oxford University Press, 1983).
12. For a recent work which aims to recast this framework, see Marya Schechtman, *The Constitution of Selves* (Ithaca, NY: Cornell University Press, 1996).
13. Parfit, "Personal Identity," *Philosophical Review* 80 (1971): pp. 3–27; see also his *Reasons and Persons* (Oxford: Clarendon Press, 1984).
14. See, for example, S. Shoemaker, *Self-Knowledge and Self-Identity* (Ithaca, NY: Cornell University Press, 1963); Williams, "The Self and the Future." For further discussion, see the articles collected in H. Noonan, ed., *Personal Identity*, International Research Library of Philosophy (Aldershot: Dartmouth Publishing Company, 1993).
15. See, for example, David Wiggins, "Locke, Butler and the Stream of Consciousness: and Men as a Natural Kind," *Philosophy* 51 (1976): pp. 131–58; Parfit, *Reasons and Persons.*
16. See, for example, Shoemaker's example in Shoemaker and Swinburne, *Personal Identity*, pp. 108–11.
17. Williams, "The Self and the Future," p. 46.
18. B. Williams, "Personal Identity and Individuation," in *Problems of the Self*, pp. 11–12.
19. Harold Noonan, *Personal Identity* (London: Routledge, 1989), p. 4.
20. A. Quinton, "The Soul," in *Personal Identity*, J. Perry, ed., (Berkeley: University of California Press, 1975), p. 60.
21. Lewis, "Survival and Identity," p. 56.
22. Noonan, *Personal Identity*, p. 13. See also Parfit, *Reasons and Persons*, p. 205.
23. Susan J. Brison, "Outliving Oneself: Trauma, Memory and Personal Identity," in *Feminists Rethink the Self*, D. Tietjens Meyers, ed., (Boulder, CO: Westview Press, 1997), pp. 12–39. Sections 4 and 5 of this chapter are deeply indebted to this article. For discussion of some closely related issues, see Sue Campbell, "Women, 'False Memory' and Personal Identity," *Hypatia* 12, no. 2 (1997): pp. 51–62.
24. See, for example, Moira Gatens, "Woman and Her Double(s)," in *Imaginary Bodies*, pp. 29–45; Grosz, *Volatile Bodies*, pp. 62–85. For a useful discussion of ways in which the notion of a body image is used, see Brian O'Shaughnessy, "Proprioception and the Body Image," in *The Body and the Self*, J. L. Bermudez, N. Eilan, and A. Marcel, eds., (Cambridge, MA: MIT Press, 1995).
25. Jacques Lacan, *Ecrits* (London: Tavistock, 1977); Grosz, *Jacques Lacan: A Feminist Introduction*; Bice Benvenuto and Roger Kennedy, *The Works of Jacques Lacan: An Introduction* (London: Free Association Books, 1986).
26. For a non-psychoanalytic treatment of this theme, see Annette Baier, "Mixing Memory and Desire," in *Postures of the Mind* (London: Methuen, 1985).
27. Charlotte Delbo, *Days and Memory*, Rosette C. Lamont, trans., (New Haven: Yale University Press, 1985). Quoted by Brison, "Outliving Oneself," p. 19.
28. Ibid., p. 21.
29. Elizabeth Grosz, *Volatile Bodies*, pp. 76–7.
30. Quinton, "The Soul," p. 64.
31. Ibid., pp. 65–6.
32. See, for example, the discussion of phantom limbs in Paul Schilder, *The Image and Appearance of the Human Body* (New York: International Universities Press, 1978), p. 64.
33. Williams, "The Self and the Future," pp. 48ff.
34. With some exceptions—see, for example, the animalist views defended in Eric Olson, *The Human Animal: Personal Identity without Psychology* (Oxford: Oxford University Press, 1997) and Paul Snowdon, "Persons, Animals and Bodies," in *The Body and the Self*, J. L. Bermudez, N. Eilan, and A. Marcel, eds., (Cambridge, MA: MIT Press, 1995)—even philosophers who do not regard psychological continuity as essential to personal identity continue to treat the body as a container for the mind.
35. See, for example, Lloyd, *The Man of Reason*; Spelman, "Woman as Body."
36. There are several variants of this view. For discussion of Freud, see Juliet Mitchell, *Psychoanalysis and Feminism*, and Jane Gallop, *Feminism and*

Psychoanalysis; on object relations theory, see Benjamin, *The Bonds of Love*; and for the view that these patterns of development are to be explained by child-rearing practices, see Chodorow, *The Reproduction of Mothering.*

37. For discussion of this view, see Le Dœuff, *The Philosophical Imaginary*; Deutscher, *Yielding Gender.*
38. At the same time, they are implicitly putting aside other dimensions of identity, for instance racial identity, which may be intimately connected to the body.
39. For a particularly helpful discussion of these instabilities, see Deutscher, *Yielding Gender.*
40. Janet Lewis, *The Wife of Martin Guerre* (Harmondsworth: Penguin, 1977). Originally published 1941.
41. Honoré de Balzac, *Le Colonel Chabert*, M. Didier, ed., (Paris: Société des textes français modernes, 1961).
42. Many helpful comments were made on an earlier draft of this essay. I am grateful to the contributors to a conference on Feminism and the Philosophy of Mind held at the University of London; to the Philosophy Department seminars at University College Dublin and at the University of York; and to John Dupré, Miranda Fricker, Jennifer Hornsby, Moira Gatens, Kathleen Lennon, Quentin Skinner, and Catherine Wilson.

Talking Identity

Kwame Anthony Appiah

Until the middle of the twentieth century, no one who was asked about a person's identity would have mentioned race, sex, class, nationality, region, or religion. When George Eliot writes in *Middlemarch* that Rosamond 'was almost losing the sense of her identity,' it's because Rosamond is faced with profoundly new experiences when she learns that Will Ladislaw, the man she thinks she loves, is hopelessly devoted to someone else.[1] Identity here is utterly particular and personal. The identities we think of today, on the other hand, are shared, often, with millions or billions of others. They are social.

One looks in vain for talk of such identities in the social science of the early twentieth century. In *Mind, Self, and Society*, published in 1934, George Herbert Mead outlined an influential theory of the self as the product of an 'I' responding to the social demands of others, which, once internalized, formed what he called the 'me.' But in that great classic of early twentieth-century social thought, you'll never find the word 'identity' used in our modern sense. Talk of identity really takes off in developmental psychology after the Second World War, with the influential work of the psychologist Erik Erikson. In his first book, *Childhood and Society*, published in 1950, he uses the term in more than one way; crucially, though, he recognizes the importance of social roles and group memberships in shaping one's sense of self, which he called, in psychoanalytic language, an 'ego identity.' Later on, Erikson explored the crises of identity in the lives of Martin Luther and Mahatma Gandhi, and published books with titles like *Identity and the Life Cycle* (1959), *Identity: Youth and Crisis* (1968), and *Dimensions of a New Identity* (1974).

Erikson, who grew up in southwest Germany, told a tale of his own origins that sits right at the heart of our contemporary notions.

> My stepfather was the only professional man (and a highly respected one) in an intensely Jewish small bourgeois family, while I (coming from a racially mixed Scandinavian background) was blond and blue-eyed, and grew flagrantly tall. Before long, then, I acquired the nickname 'goy' in my stepfather's temple; while to my schoolmates, I was a 'Jew.'

I'm guessing that, while his Jewish confreres did use the Yiddish word for a gentile, those German kids didn't always use a word as polite as 'Jew.' His biological father had been a Dane named Salomonsen; his adopted father's name was Homburger. But at some point he took the last name of Erikson, which, as his daughter once observed drily, suggested that he was father to himself. In a sense, then, he was his own creation.[2] Identity, we can surely conclude, was a fraught issue for him personally.

In his first book, Erikson offered a theory as to why, as he put it, 'we'—and given our subject, it's worth noticing that he seems to mean 'we Americans'—'began to conceptualize questions of identity.' He thought that identity had become a problem in the United States because the country was 'trying to make a super-identity out of all the identities imported by its constituent immigrants'; and, he continued, 'we do so at a time when rapidly increasing mechanization threatens these essentially agrarian and patrician identities in their lands of origin.'[3] It's a good story. But I do not believe it. As we shall see throughout this book, identity, in our sense, was a problem long before we began to talk about it in this modern way.

If Erikson, weaving between personal and collective forms of identity, gave the term broad currency, the influential American sociologist Alvin W. Gouldner was among the first to offer a detailed definition of social identity as such. 'It seems that what is meant by a 'position' is the social identity which has been assigned to a person by members of his group,' he wrote in a 1957 essay. And he proposed an account of what this means, practically, in social life. First, he thought, people 'observe or impute to a person certain characteristics,' which allows them to 'answer the question 'Who is he?'' Next, 'these observed or imputed characteristics are . . . interpreted in terms of a set of culturally prescribed *categories*.'

> In this manner the individual is 'pigeonholed'; that is, he is held to be a certain 'type' of person, a teacher, Negro, boy, man, or woman. The process by which the individual is classified by others in his group, in terms of the culturally prescribed categories, can be called the assignment of a 'social identity.' The types or categories to which he has been assigned *are* his social identities. . . . Corresponding to different social identities are differing sets of expectations, differing configurations of rights and obligations.[4]

As you'll see, I think that Gouldner got a lot right.

Appeals to identity swelled through the sixties and, by the end of the seventies, many societies had political movements grounded in gender and sexuality, race, religion, and ethnicity (even as class politics frequently receded into the background). In more than a few places, regionally based movements that sought to undo often long-established states spoke the language of national identity. In Europe alone, there's Scottish, Welsh, Catalan, Basque, Padanian, and Flemish nationalism; near the end of the twentieth century, Yugoslavia collapsed into a collection of distinct countries; there are rumblings in Brittany, Corsica, and Normandy . . . and that's far from a complete list.

A Little Theory

I have been writing and ruminating on questions of identity for more than three decades now. My theoretical thinking about identity began, actually, with thoughts about race, because I was genuinely puzzled by the different ways in which people in different places responded to my appearance. That wasn't so much the case in Asante, where, so it seemed to me, one local parent was usually enough to belong. Jerry Rawlings, Ghana's head of state from 1981 to 2001, had a father from Scotland; he wasn't chosen by the people originally—he came to power twice through coups d'état—but his fellow countrymen eventually elected him to the presidency twice. Unlike my three sisters, born, like my father, in Asante, I have never been a Ghanaian citizen. I was born in England, before Ghana's independence, with an English mother, and showed up in Asante at the age of one. So I'd have had to apply for Ghanaian citizenship, and my parents never applied for me. By the time it was up to me, I was used to being a Ghanaian with a British passport. My father, as president of the Ghana Bar Association, was once involved in writing one of our many constitutions. 'Why don't you change the rules, so that I can be both Ghanaian and British?' I asked him. 'Citizenship,' he told me, 'is unitary.' I could see I wasn't going to get anywhere with him! But, despite my lack of that legal connection, sometimes, when I do something noteworthy, I am claimed, at least by some, for the place that is home to half my ancestry.

The story in England was complex, too. In my grandmother's village, Minchinhampton, in Gloucestershire, where I spent much time in my childhood, those we knew never appeared to doubt our right to be there. My aunt and uncle lived in this picturesque market town in the west of England, too. My aunt had been born there. My grandfather had spent time as a child at a house in the valley, which belonged to his uncle, whose mill had once woven cloth for the tunics of British soldiers and green baize for billiard tables. My great-grandfather, Alfred Cripps, had briefly served as the member of parliament for Stroud, a few miles to the north, and *his* great-grandfather, Joseph Cripps, had

represented Cirencester, a few more miles east, for much of the first half of the nineteenth century. And there were Crippses in that area—some buried in Cirencester churchyard—dating back to the seventeenth century.

But the skins and the African ancestry I shared with my sisters marked us out as different, in ways we weren't always conscious of. I recall going to a sports day, a few decades ago, at a school in Dorset I'd attended as a preteen, and coming upon an elderly man who had been headmaster in my day. 'You won't remember me,' I apologized, as I introduced myself to him. Hearing my name, he brightened and took my hand warmly. 'Of course I remember you,' he said. 'You were our first colored head boy.' When I was young, the idea that you could be properly English and not white seemed fairly uncommon. Even in the first decade of the twenty-first century, I remember the puzzled response of an older Englishwoman who had just heard a paper on race I gave at the Aristotelian Society in London. 'She just didn't understand how I could really be English. And no talk of thirteenth-century ancestors in Oxfordshire could persuade her!' In America, once I got there, things seemed at first relatively simple. I had an African father and so, like President Obama later, I was black. But the story here, too, is complicated . . . and has changed over the years, in part because of the rise of the idea of mixed-race people as an identity group. Color and citizenship, however, were quite separate matters: after the Civil War no sensible person doubted you could be black and American, at least so far as the law was concerned, despite a persistent undercurrent of white racial nationalism. I'll say more about the ideas of race that shaped these experiences later, but I hope it's clear why I might have ended up puzzled about how to make sense of them.

When I turned, over the years, to thinking about nationality and class and culture and religion as sources of identity, and added in gender and sexual orientation, I began to see three ways in which these very disparate ways of grouping people do have some important things in common.

Labels and Why They Matter

The first is obvious: every identity comes with labels, so understanding identities requires first that you have some idea about how to apply them. Explaining to someone what Ewes or Jains or *kothis* are begins with some suggestion as to what it is about people that makes each label appropriate for them. That way, you could look for someone of that identity, or try to decide, of someone you'd met, whether the label applied.

So, the label 'Ewe' (usually pronounced eh-vey or eh-wey) is an ethnic label, what social scientists call an 'ethnonym'; which means that if your parents are both Ewe, you're Ewe, too. It applies, in the first place, to people who speak one of the many dialects of a language that is called 'Ewe,' most of whom live in Ghana or Togo, though there are some in many other parts of West Africa and, increasingly, around the world. As is typical of ethnic labels, there can be arguments about whether it applies to someone. If only one of your parents is Ewe, and you never learned any of the many dialects of the Ewe language, are you Ewe? Does it matter (given that the Ewe are patrilineal) if the parent was your mother rather than your father? And, since Ewe belongs to a larger group of languages (usually called 'Gbe' because that's the word for language in all of them) that shade off into one another, it's not easy to say exactly where the boundaries between Ewe people and other Gbe-speaking people lie. (Imagine looking for the boundaries of Southern speech in America or a cockney accent in London and you'll grasp the difficulty.) Nevertheless, large numbers of people in Ghana and Togo will claim that they're Ewe and many of their neighbors will agree.

That's because of the second important thing identities share: they matter to people. And they matter, first, because having an identity can give you a sense of how you fit into the social world. Every identity makes it possible, that is, for you to speak as one 'I' among some 'us': to belong to some 'we.' But a further crucial aspect of what identities offer is that they give you reasons for doing things. That's true about being a Jain, which means you belong to a particular Indian religious tradition. Most Jains are the children of two Jains (just as most Ewes are the children of two Ewes), but there's much more to it than that. And anyone can join who is willing to follow the path set by the *jinas*, souls who have been liberated by conquering their passions and can spend a blissful eternity at the summit of the universe. Jains are typically expected to heed five *vratas*, which are vows or forms of devotion. These are: nonviolence, not lying, not stealing, chastity, and

nonpossessiveness. (Like taboos, which are also central to many identities, the *vratas* define who you are by *what* as well as *who* you are *not*. There's a lot of 'Thou shalt not's' in the Ten Commandments, too.)[5]

The detailed content of each of these ideals depends, among other things, on whether you are a layperson on the one hand, or a monk or nun on the other. The general point, though, is that there are things people do and don't do *because they are Jains*. By this, I mean only that they themselves think from time to time, 'I should be faithful to my spouse . . . or speak the truth . . . or avoid harming this animal . . . because I am a Jain.' They do that, in part, because they know they live in a world where not everyone is a Jain, and that other people with other religions may have different ideas about how to behave.

Though there are Ewe religious traditions (lots of different ones), being Ewe isn't, by contrast, a religious identity, and doesn't come with the same sort of specified ethical codes. Ewes can be Muslim, Protestant, or Catholic, and many practice the traditional rites that go by the name of voodoo. (Like the Haitians, they borrowed this word from the Fon peoples, who are their neighbors. It means 'spirit.') But, all the same, Ewe people sometimes say to themselves, 'As an Ewe, I should . . .' and go on to specify something they believe they should do or refrain from doing. They do things, in short, because they are Ewe. And this, too, depends, in part, on their recognition that not everyone is Ewe, and that non-Ewes may well behave differently.

People who give reasons like these—'Because I'm a this, I should do that'—are not just accepting the fact that the label applies to them; they are giving what a philosopher would call 'normative significance' to their membership in that group. They're saying that the identity matters for practical life: for their emotions and their deeds. And one of the commonest ways in which it matters is that they feel some sort of solidarity with other members of the group. Their common identity gives them reason, they think, to care about and help one another. It creates what you could call norms of identification: rules about how you should behave, given your identity.

But just as there's usually contest or conflict about the boundaries of the group, about who's in and who's out, there's almost always disagreement about what normative significance an identity has. How much can one Ewe or one Jain legitimately ask of another? Does being Ewe mean you ought to teach the Ewe language to your children? Most Jains think that their religion requires them to be vegetarian, but not all agree that you must also avoid milk products. And so on. While each Ewe or each Jain will have done things because of their identity, they won't always do the same things. Still, because these identities sometimes help them answer the question 'What should I do?' they're important in shaping their everyday lives.

One further reason that's true is the third feature all identities share: not only does your identity give *you* reasons to do things, it can give others reasons to do things *to* you. I've already mentioned something people can do to you because of your identity: they can help you just because you share an identity with them. But among the most significant things people do with identities is use them as the basis of hierarchies of status and respect and of structures of power. Caste in South Asia means some people are born into a higher status than others—as Brahmins, for example. These are members of the priestly caste, who are 'polluted' by contact with members of castes that are regarded as lower. In many places in the world one ethnic or racial group regards its members as superior to others, and assumes the right to better treatment. The English poet Shelley, in 'Ozymandias,' refers to the 'frown / and wrinkled lip and sneer of cold command' on the stone face of the sculpture of a long-dead Pharaoh. The royal ancestry of this 'king of kings' would have meant that he was used to obedience. Dominant identities can mean that people will treat you as a source of authority; subordinate identities can mean you and your interests will be trampled upon or ignored.

And so an important form of struggle over identity occurs when people challenge the assumptions that lead to unequal distributions of power. The world is full of burdensome identities, whose price is that other people treat you with disrespect. *Kothis* in India know this very well. They are people who, though assigned a male identity at birth, themselves identify as feminine, and experience erotic attraction to men who are more typically masculine. And *kothis* have been subjected over the years to insult and abuse, and to rejection by their families; many of them have been forced by their marginal position into sex work. In recent years, emerging ideas about gender and sexuality—about homosexuality, intersexuality, and transgender identity, and about the complexity of the

connection between biological sex and human behavior—have created movements that seek to alleviate the social exclusion of people whose gender and sexuality fall outside traditional norms. The Indian Supreme Court has even declared that individuals are entitled to be recognized as male, female, or third-gender, as they themselves decide.

Once identities exist, people tend to form a picture of a typical member of the group. Stereotypes develop. They may have more or less foundation in reality, but they are almost always critically wrong about something. *Kothis*, some Indians think, really want to be women: they are, many people suppose, what Europeans and Americans would now often call 'transsexual.' But that's not necessarily so. Ewes, other Ghanaians fear, are particularly likely to use 'juju'—witchcraft or 'black magic'—against their enemies. But witchcraft is traditional all over Ghana, so this isn't, actually, much of a distinction. (I once wrote an account of my father's funeral, in the course of which I discussed how we had to deal with the threat of witchcraft in our family. We, as you know, were Asante, not Ewe.)[6] People believe that Jains are so obsessed with nonviolence that they insist on covering their faces with white cloth to avoid killing insects by ingesting them. In fact, most Jains don't wear the *muhapatti*, as the white cloth is called, and its use has a variety of rationales that have nothing to do with saving the lives of insects.

In sum, identities come, first, with labels and ideas about why and to whom they should be applied. Second, your identity shapes your thoughts about how you should behave; and, third, it affects the way other people treat you. Finally, all these dimensions of identity are contestable, always up for dispute: who's in, what they're like, how they should behave and be treated.

Woman, Man, Other?

This picture of identity is, in effect, a generalization of ways of thinking about gender that have been pioneered by feminist scholars. Feminism made use of theoretical ideas in the pursuit of women's equality and liberation from old patterns of oppression. All human societies have some form of gender system—some way of thinking about the significance of the distinction between men and women. But feminist theories allow us to see what all the multitudinous systems of gender have in common while, at the same time, allowing us to keep track of their differences.

Let's rehearse some details. The vast majority of human bodies can be recognized as belonging to one of two biological kinds. Simply examining the genitalia—the organs of sexual reproduction—will usually allow you to see that someone is biologically male (because he has a penis, scrotum, and testicles), or biologically female (because she has a vagina, labia, uterus, and ovaries). In adults, you may be able to make the discrimination at a glance: the breasts of the biological females will grow at puberty, facial hair will develop in the males, their voices will deepen, and so on. Chromosomal analysis will also usually allow you to discover that the males have an X and a Y chromosome and the females two X's. Knowing all this, ordinary people and medical experts alike can apply the labels 'woman' and 'man.'

But these turn out to be only two of the great variety of regularly occurring combinations of sex chromosomes and sexual morphology. In the standard case, the sex organs of human males and females initially develop in the same way in the embryo, and in the early stages the structure that will eventually become either an ovary or a testis is called the 'indifferent gonad.' In the typical male fetus, genes on the Y chromosome trigger changes that produce the male testis, and thus the production of hormones that influence the development of other sex-related structures. Absent this stimulus, the indifferent gonad turns into an ovary. It's the presence of the Y chromosome, then, that makes you a male.

That's the basic story. But there are many variations. One possibility is that, despite the presence of a Y chromosome, female external genitalia emerge. This can happen for a variety of reasons, one of which is androgen insensitivity syndrome (AIS), which means that your cells are not normally sensitive to male sex hormones. XY people with AIS can have either male or female external genitalia, or something in between, but the females aren't fertile because they have testes in place of ovaries.

There are other ways in which a mismatch between external appearance and your sex chromosomes can develop. Maternal androgens can turn the genitalia in the male direction, producing someone who is XX but externally male. So a fertilized human egg that is clearly XY can end up producing someone who looks

like a woman and one that is XX can produce someone who looks like a man. And there are various other possible combinations: penis and ovaries, vagina and abdominal testes, external genitalia that are intermediate, and so on.

And that's all assuming you start out with two sex chromosomes. In fact, there are some people who are XO, having just one X chromosome. This is Turner syndrome, and people who have it have the bodies of women, though they're usually infertile and often shorter than average. (You need at least one X chromosome to survive—the Y chromosome is much smaller than the X and lacks some of the genes on the X that are essential for human life—which is why there are no OY males.) People with Turner syndrome sometimes have medical problems; but among the best-known people with the condition are a world champion gymnast, Missy Marlowe, who has been a spokesperson for the Turner Syndrome Society, and the Oscar-winning actress Linda Hunt.

Then there are people with an extra X chromosome—XXY or XXX—and, rarely, even more. Because in normal female cells only one of the X chromosomes is active (the other existing in a contracted and largely inactive form called a 'Barr body'), these extra X's don't usually make a huge difference: if you have a Y chromosome, you'll look male; if you don't, you'll look female. While all these variations are rare, they do mean that even at the level of physical morphology, there just isn't a sharp division of human beings into two sexes.

All societies start with this spectrum of morphological possibilities. They are a basic part of our human biology. Because the intermediate cases are statistically rare, many people in smaller communities may never interact with anyone but XY males and XX females, with a sexual morphology in the standard range. Given this variability, it's not surprising that different societies have come up with different ways of assigning people to a gender. In many places, surgeons have often tried to 'tidy up' the genitalia of babies born with nonstandard sexual bodies, soon after birth. So they've sought to bring everybody into a binary system, in which everyone is more or less clearly male or female. Not everyone agrees that this is a good idea.

In India, *kothis* have long been treated as neither men nor women; and *kothi* interacts with another form of South Asian gender identity, whose label is *hijra*.[7] *Hijras*, as a committee of Indian government experts put it in 2014, 'are biological males who reject their 'masculine' identity in due course of time to identify either as women, or 'not-men,' or 'in-between man and woman,' or 'neither man nor woman."[8] But *hijras* have a long tradition of living as a community with rites of entry, dressing in women's clothes, and wearing women's makeup. *Kothis*, on the other hand, generally cross-dress only in private or when socializing with one another. Many don't cross-dress at all. Sometimes *hijras* have sought gender-reassignment surgery; in the past, many underwent castration. Notice that neither of these terms corresponds to our terms 'transgender' or 'homosexual,' since (to mention only one difference) the South Asian categories don't cover what we would call either F-to-M transsexuals or lesbians.

Anjum, one of the protagonists of Arundhati Roy's extravagantly rambling novel *The Ministry of Utmost Happiness*, is what would once have been called a hermaphrodite: she is raised as a boy named Aftab because her mother seeks to conceal the fact that she has both male and female sexual organs.[9] But the boy, Aftab, doesn't want to be a boy, though he doesn't yet know what he *does* want to be. And then, one spring morning,

> Aftab saw a tall, slim-hipped woman, wearing bright lipstick, gold high heels and a shiny green satin salwar kameez, buying bangles from Mir the eagle-seller. . . . Aftab had never seen anyone like the tall woman with the lipstick. . . . He wanted to be her.[10]

Aftab follows this colorful *hijra* home to the Khwabgah—the town house where the *hijras* of her part of Delhi live—and finds there a whole community of people to whom she somehow knows she belongs. Being *hijra* is more than having a male body and feminine style: as we learn through the novel, *hijras* have a role in Indian life, and so identifying as one entails more than just dressing up. I am relaying the account of a fictional character; but Anjum, I'm told, is based on a real person.

On the other side of the globe, too, some of the Indian tribes of North America once recognized a variety of genders. The Navajo, in the nineteenth century, for example, called intersexes who were masculine *dilbaa*, and those who were feminine *nádleehí*.[11] They took up special roles in religious life. More recently, many American Indian activists have come to use the neologism 'Two Spirit' to speak of those who do not fit easily into the categories of man and woman, in one way or another. The term reflects the fact that people who were neither men

nor women, but had something—a spirit—from each, played special religious roles in many American Indian societies. And this is how a lot of contemporary American Indians, whom many other Americans would call lesbian, gay, or trans, now choose to identify themselves.

What feminist theorists taught us to see was that when we speak of men and women, or of other genders, we're not talking just about bodies. In calling a child a girl or a boy—in applying that label—every society is assuming more than that the child has a certain sexual morphology. And so we distinguish now between sex (the biological situation) and gender (the whole set of ideas about what women and men will be like and about how they should behave). Some researchers have argued that one out of every hundred children is intersex in some way.[12] In a world of more than 7 billion people, 1 percent of the population is a whole lot of individuals. So midwives and obstetricians and others who witness many births may well come across such cases from time to time and have to decide what (if anything) they should do about them. But even in a world of XX females and XY males, gender would impose a great deal of structure on thinking about what women and men are, or should be, like.

Why? Because identities, as I said, involve labels and stereotypes. That is obvious in the case of gender. If you're labeled a man, in most societies, you are supposed to be sexually interested in women, to walk and use your hands in a 'manly' way, to be more physically aggressive than women, and so on. Women should be sexually interested in men, walk and talk in a feminine way, be gentler than men, and all the rest. I have been using the words 'male' and 'female' to talk about bodily differences: but we need words to mark these other forms of difference built upon that foundation. So I'm going to continue using 'masculine' and 'feminine' to talk about the forms of thought, feeling, and behavior that our pictures of gender lead us to expect in men and in women, respectively. Men are—and are supposed to be—well, masculine. Men should lead, women should follow; women obey, men command. And that 'supposed to be' and those 'shoulds' are both descriptive (this is what we expect men and women to be like) and normative (this is what we think is right). But, once more, people disagree about these traditional claims about what men and women—and people who think they are neither—should be like. And these notions clearly vary across time and space; many contemporary New Yorkers will assume that a woman might be tough as nails and that a man could be, in Shakespeare's phrase, 'as mild and gentle as the cradle-babe.'

Labels, stereotypes, and ideas about how you should behave: these, I said, are there in every identity. And gender has the last of the traits I mentioned, too; it involves ideas not just about how you should behave, but also about how others should behave toward you. In the old days, there was a gentlemanly code (reflecting hierarchies of power) of opening doors and holding out chairs and paying for meals, and such. New norms of treatment have emerged, some relating to how women interact with one another, or how men interact with men, some relating to how women and men interact. Next time you're in a crowded elevator in a modern cosmopolitan city, watch to see whether the men stay back to let the women exit first. Now imagine the life of a woman who insists, in the name of challenging older stereotypes, on refusing such offers. Identities, in this way, can be said to have both a subjective dimension and an objective one: an identity cannot simply be imposed upon me, willy-nilly, but neither is an identity simply up to me, a contrivance that I can shape however I please.

The shape of one identity can also be contoured by your other identities. To be an Ewe woman is not just a matter of being a woman *and* being an Ewe, in some easy act of addition. An Ewe woman faces certain expectations—expectations to meet and expectations to be met—that are peculiar to Ewe womanhood. To be Chinese and gay means something different if you're a native of San Francisco than it does if you're a native of Zhumadian, in China's Henan Province, where, not long ago, a hospital institutionalized a man for 'sexual preference disorder' and forced him to undergo conversion therapy. The social import of an identity can vary with wealth, age, disability, weight, employment status, and any other social coordinate you might think of. In political contexts, though, an identity group can be avowedly global ('Workers of the world, unite!'; 'Women of the world, rise up!'), sometimes with older forms of identity melded into larger, newer composite ones (people of color; LGBTQ). Identity is here enlisted for purposes of solidarity. To be sure, being a member of an identity group that is, in certain respects, subordinated doesn't necessarily make you sympathetic toward another (black Americans, often for religious reasons, are more likely to oppose same-sex marriage than whites), and sometimes

the fiercest antipathy toward an identity group (as with squabbles among religious sects over 'heresies') emanates from an intimately adjacent one.

These complex interactions between identities—which we see in the case of *kothis*, say, where ideas of sexuality and gender both matter—are one reason that Kimberlé Crenshaw, a feminist legal theorist and civil rights activist, introduced the idea of *intersectionality*. She wanted to talk about the ways in which our many identities interact to produce effects that are not simply the sum of each of them. Being a black lesbian is not a matter of combining African-American, female, and homosexual norms of identification: LGBTQ norms of identification can depend on your race and your gender. Nor are the negative social responses to black lesbians simply a combination of the racist and homophobic responses that also affect black gay men and the sexist responses experienced by middle-class white women.[13] Racism can make white men fear black men and abuse black women. Homophobia can lead men in South Africa to rape gay women but murder gay men. Sexism in the 1950s kept middle-class white women at home and sent working-class black women to work for them. Examples of intersectionality proliferate.

The fact of intersectionality raises a problem for one of the ways people bring their identities to bear nowadays. Say that Joe, who's a white man, claims to speak *as a man*, or *as a white person*. What does that mean, beyond the fact that he's speaking *and* he's male or white? Having an identity doesn't, by itself, authorize you to speak on behalf of everyone of that identity. The privilege of representing a group has to be granted somehow. So, absent evidence that he's somehow been given or otherwise earned the authority, it can't mean that Joe is speaking for all white people or for all men. You might think that he has at least the authority of experience to speak about what it's like to live as a white man. Is that something that a white man can discuss with special knowledge, just because he's been through it? Not if we take the point about intersectionality. For, to the extent that how people treat you affects your experience, intersectionality makes it likely that there will be differences in the experience of, say, gay white men and straight white men; and, if Joe had grown up in Northern Ireland, as a gay white Catholic man, his gay white Protestant male friends might well have rather different experiences, too. And, once you think about it a little, you can see that, while your identity affects your experience, there's no guarantee that what you've learned from it is going to be the same as what other people of the same identity have learned.

Yet the familiar fact that our identities are multiple and can interact in complicated ways is consistent with a pretty frugal account of what, conceptually, any identity consists in: taking a label and a picture of how to apply it that entrains norms about how people who have the label should behave and how they should be treated.

Habitus

None of that is new, of course. 'Woman,' 'Ewe,' 'Jain,' '*kothi*,' '*hijra*,' were like that long before scholars started talking about social identities. From Shakespeare to Gilbert and Sullivan, there's a long history of pride in being English that echoes portentously from Henry V's speech at Harfleur addressing his troops as 'good yeoman, whose limbs were made in England,' to the more comical strains of *H.M.S. Pinafore*, where the Boatswain affirms that Ralph, the humble cabin boy,

> . . . has said it
> And it's greatly to his credit,
> That he is an Englishman.

As a teenager, I delighted in a satiric recording by Michael Flanders and Donald Swann who insisted, 'the English are best,' and sang cheerfully that they 'wouldn't give tuppence for all of the rest!' What's new is thinking of these diverse sorts of labels—Englishman, woman, *kothi*, and so on—as things of the same kind. The rise of identity is the rise of that thought.

Once you think that thought, you can ask questions about the social and psychological significance of identities. And a great deal of modern psychology and sociology has been about just that. To complete my sketch of a theory of identity, I want to point to three important discoveries that have emerged in the course of such research.

The first is about how central identity is to the way we deploy our bodies. The French sociologist Pierre Bourdieu put it this way. Each of us has what he called a *habitus*: a set of dispositions to respond more or less spontaneously to the world in particular ways, without much thought. Your habitus is trained into you starting from childhood. Parents tell you not to

speak with your mouth full, to sit up straight, not to touch your food with your left hand, and so on, and thus form table manners that are likely to stick with you all your life.[14] Once they are inculcated, these habits aren't consciously associated with an identity: middle-class English people don't consciously decide to hold their knives in their right hands in order to act English, any more than Ghanaians use only their right hands to eat in order to display that they're Ghanaian. But these habits were nevertheless shaped by their identities.

Bourdieu held a prestigious chair at Paris's Collège de France and had a career in the heart of the French academic elite, but he grew up in a village in southwest France, the son of a farmer turned postman, and retained a critical distance from the social codes that surrounded him as an adult. He saw the habitus as grounded in the distinctive way in which a person used his or her body, what he called the 'bodily hexis,' 'a durable way of standing, speaking, walking, and thereby of feeling and thinking.'[15] (I wouldn't bother to introduce this horrible jargon if it weren't going to be useful later!) But it also includes modes of speech, such as what he once called the French 'intellectuals' new style of speech—a little hesitant, even mumbling, interrogative ('*non*?') and faltering,' that had replaced 'the old professorial style (with its long sentences, imperfect subjunctives, etc.)'[16]

You learn how to dress as a man or as a woman in ways that are shaped by the clothes you are given as a child, which themselves are selected because of your gender. You learn how to walk, in part, by watching other boys or girls walking. If a man wears makeup—as the Prophet Muhammad wore kohl around his eyes and Maasai men paint their faces with ochre—he'll wear it in the way other men do; in most societies, women wear different styles of makeup from men. But none of this is particularly conscious: when I buy a jacket, I'm not thinking, 'Must dress like a man.' When I walk, I don't consciously reflect that I'm not walking like a woman. Nevertheless, my clothes and my gait reflect my gender and the models of masculinity I have relied on. As Aftab/Anjum reminds us, it is through our identifications that we recognize our models.

Gender norms are enforced in myriad ways. I recall a lesson delivered, when I was eight or nine, by the decidedly old-fashioned headmaster of a school I then attended. His name was Reverend Hankey (you can imagine what a group of prepubescent boys made of that) and one day he gave us a stern lecture that there was to be no roughhousing—'ragging' he called it, in the argot of the day—in the combination room where we hung out between classes. A few days later, he came into the room and found me sitting on the chest of a fellow student, who, if memory serves me right, we called 'Piggy,' because his family name was Hogsflesh. I was tickling Piggy as he struggled to escape. We were summoned to the headmaster's office, where my fellow ragger went first. I heard the sound of four whacks on his bottom with a bamboo cane. So I girded myself up for the same. But after the third whack, Reverend Hankey stopped. 'I'm giving you one less than Hogsflesh,' he pronounced solemnly. 'I said no ragging. But if you are going to rag, it's better to be on top.' (The school, if not Reverend Hankey's ideals of masculinity, dissolved a few years later.)

Girls in Japan see other girls covering their mouths when they laugh. They do likewise. If they don't, they are corrected. But because of this, some gay men in Japan also cover their mouths when they laugh, and this reflects the fact that they identify to some degree with women. Because the ways in which men and women dress and walk in different social groups are different, you end up walking and dressing in ways that reflect your identity, not just your gender but your class and your ethnicity. The swagger of some inner-city African-American men is as much a reflection of class as of race and gender. The *Encyclopedia of African American Popular Culture* meticulously describes the style of the 'pimp walk' as a 'demonstration of cool masculinity . . . a cocksure combination of leisurely strutting, black aesthetics and public performance . . .'[17] A woman who walked that way would strike others as strange; and most patients would be skeptical of doctors who walked that way, whatever their race or social origins.

Among the most significant elements of your bodily hexis, Bourdieu thought, were habits of using your mouth; people acquire a distinctive accent, a recognizable way of speaking, that reflects dimensions of their social identity.[18] An accent can distinguish a class or even a profession, as does the speech of the ludicrous cavalry officer, Wellesley Ponto, in Thackeray's 1848 *Book of Snobs*. Thackeray describes him as 'a gaunt and bony youth,' who explains frankly why he needs his father to pay off the debts he has acquired living up to the style of his more prosperous fellow officers.

> 'Gad!' says he, 'our wedgment's so doothid exthpenthif. Must hunt, you know. A man couldn't live in the wedgment if he didn't. Mess expenses enawmuth.'

Thackeray was satirizing such people, but the accent was real. And the drawl and the lisp were meant to express an aristocratic indifference to haste and a languorous unwillingness to waste energy in conversation. The stiff upper lip here was not just figurative. Bourdieu in his work offered another example of a connection between the overall habitus of one class and another, in a fascinating discussion of the distinction between two words in French that can both refer to the mouth, *bouche* and *gueule*. The sociologist John Thompson summarizes the analysis very nicely:

> In French there is a distinction between a closed, pinched mouth (*la bouche*) and a large open mouth (*la gueule*). Individuals from working-class backgrounds tend to draw a socially and sexually overdetermined opposition between these terms: *la bouche* is associated with the bourgeois and the feminine (e.g., tight-lipped), whereas *la gueule* is associated with the popular and the masculine (e.g., 'big mouth,' 'loud mouth').

As a result, speaking like a bourgeois can seem to a working-class Frenchman to betray his masculinity.[19]

Most of us do not think of our accent as something we consciously chose, nor do we usually reflect upon the fact that the reason we speak the way we do reflects many dimensions of identity beyond the region and the class we come from. Our accent is part of our habitus, one of the routine ways in which we use our bodies. I mentioned in the introduction that my English accent has sometimes puzzled taxi drivers, in part because they're not used to brown-skinned people sounding like members of the English upper middle classes; but, like most people, I speak in the way my school friends spoke when I was growing up. It's unusual to acquire the fluency of a native speaker in a new language as an adult. But a Ghanaian man I know, who has lived in Japan for a long time, told me that he once approached a Japanese woman who was having trouble with a bicycle with a flat tire. When he first started speaking to her, she didn't look up. His Japanese sounded quite normal to her. When she finally glanced up at him, he could see a look of astonishment cross her face. She hadn't expected to see a black-skinned foreigner. (For the record, the story turns out well: she's now his wife.)

Habitus and identity are connected by the fact that we recognize certain forms of behavior—accents, but also ways of walking, styles of dress—as the signs of certain forms of identity and that our identities shape our habitus unconsciously. I've said that identities matter because they give us reasons to do things, reasons we think about consciously. But the connection between identity and habitus means that identities matter in unreflective ways as well. The distinguished American social psychologist Claude Steele describes how a young black graduate student at the University of Chicago, troubled by the fearful responses of white people, takes to whistling Vivaldi as he walks down the street. The student signals his knowledge of 'high culture,' and white people (who might not know it's Vivaldi), recognize this is classical music. 'While hardly being aware of it,' Steele writes, they drop 'the stereotype of violence-proneness. . . . Fear fades from their demeanor.'[20] Sociolinguists have catalogued the many ways in which people adjust their verbal style in response to the social identities of people they're talking to, again while hardly being aware of it. I've been told that I adjust my accent in an American direction when I'm telling New York taxi drivers where I want to go. It doesn't matter that I apparently have a terrible American accent. I'm trying—without consciously meaning to—to make myself easier to understand for people who are often, like me, immigrants, and are working to understand the local dialect of English.

Essentialism

The second important psychological truth also comes with a fancy name: essentialism. Psychological research has revealed that, long before anyone instructs children to group people into categories, they're programmed to do it anyway. By the age of two, children distinguish between males and females and expect them to behave differently. And once they classify people, they behave as if each person in the group shares some inner something—an essence—that explains why they all have so much in common. 'Essentialism is the view that certain categories have an underlying reality or true nature that one cannot observe directly,' the developmental psychologist Susan Gelman says, 'but that gives an object its identity, and is responsible for other similarities that category members share.'[21] Children everywhere are

full-fledged essentialists by the time they are four to six years old.

It's not that they don't notice the superficial, visible features of people. Far from it. The color of hair and skin and other aspects of physical appearance play a role in determining what sorts of people are grouped together. I know of a distinguished black New York literary agent who finds children in the elevator of her building reaching out to her for a hug: in their world black women are nannies, and nannies are there for the hugging. What essentialism means is that children assume that these superficial differences—the ones that lead to applying the label—reflect deeper, inward differences that explain a great deal of how people behave.

Research with young children suggests that one of our most basic strategies for making sense of the world is to form the sorts of generalizations that linguists call 'generics'—generalizations like 'Tigers eat people,' and 'Women are gentle.' It also turns out that it's very hard to say what makes generics true. They're not equivalent to universal claims like 'All tigers eat people.' After all, most tigers have not eaten a person; in fact, very, very few have. As for whether women are gentle: well, which women? Certainly not the fierce Amazon regiment (whom the Fon charmingly called 'our mothers,') that served the nineteenth-century kings of Dahomey. So the generic claim that 'Women are gentle' doesn't mean *all* women are gentle; and 'Tigers eat people' doesn't mean that *most* tigers eat people. In fact, as my friend the philosopher Sarah-Jane Leslie has pointed out, an epidemiologist can sincerely say, 'Mosquitoes carry the West Nile virus,' while knowing that 99 percent of them *don't* carry it.

Generics work by a basic kind of association of ideas. So thinking 'Tigers eat people' means that, faced with a tiger, your default response is going to be to think about its eating someone—perhaps even you. 'Mosquitoes carry the West Nile virus' will have your doctor checking your temperature when she sees your mosquito bites.[22] As these examples suggest, one thing that makes it more likely that we'll accept a generic is if the property it mentions is one that we have a reason to be concerned about, like people-eating or pathogen-spreading.

But it also helps if we think of the class (tigers, women, mosquitoes) as a kind, as a group of organisms with a shared essence. And getting children to think of a group of *people* as a kind is very easy. The psychologist Marjorie Rhodes and her colleagues did the following simple experiment. They showed four-year-olds pictures of a fictional kind of person they called a Zarpie. The pictures were male and female, black, white, Latino, and Asian, young and old. With one group of kids, the experimenters made lots of generic remarks about these imaginary people—'Zarpies are scared of ladybugs,' and the like. With another group of kids, they avoided generics. ('Look at this Zarpie! He's afraid of ladybugs!') A couple of days later they showed the kids a Zarpie and said that he made a buzzing sound. It turned out that the kids who'd heard a lot of generics about Zarpies were much more likely to believe that *all* Zarpies made buzzing sounds. Generic talk encouraged them to think of Zarpies as a kind of person. And once kids think of Zarpies as a kind of person, they're more likely to infer that the behavior of one Zarpie reflects the nature of all Zarpies, that buzzing flows from the Zarpie essence.[23]

Let's put the lessons of the last two paragraphs together. I can get you to think of people—even a group of diverse-looking people of both sexes and all ages—as a kind, by making generic remarks about them. And you're more likely to accept a generic claim about a group if what it says is something negative or worrying. We humans are more likely, then, to essentialize groups about which we have negative thoughts; and more likely to have negative thoughts about groups we've essentialized. There's an unfortunate vicious circle for you. (The next time someone tells you that 'Muslims are terrorists,' you might want to bear that in mind.)

The plain fact is that we're really good at conjuring up Zarpies, and viewing them with suspicion. Take the Cagots, of the French and Spanish Pyrenees. Though they largely melted away in the nineteenth century, through migration and assimilation, the Cagots were, for a millennium, treated as pariahs, relegated to disfavored districts, even forced to use separate doors in churches, where they received the Communion wafer at the end of a stick. Because contact with the Cagots was contaminating, they were severely punished for drinking from the same water basin as others, for farming, or even for walking barefoot on the streets. Songs about them—one, recorded in the mid-nineteenth century, goes: 'Down with the Cagots, / Let's destroy them all! / Let's destroy the Cagots, / And down with them all!'—made it clear how you should regard them, but didn't tell you why. What distinguished them from their neighbors? Not their appearance. (That's why they were forced to identify themselves with badges

pinned to their clothing, often duck or goose feet, or fabric facsimiles.) Not their family names. Not their language. Not their religion. The real mystery of the Cagots, Graham Robb concludes in his history of France, 'was the fact that they had no distinguishing features at all.'[24]

In large ways and small, essentialism shapes our public history, and it shapes our personal histories as well. It's there in the responses of some white people to Claude Steele's black graduate student on the streets around the University of Chicago. It's evident, too, in the ways we are prone to assume, in the domain of gender, that 'boys will be boys' and men, men. We expect all kinds of gendered behavior in ways that suggest that there is an inner something that not only explains why (as we might imagine) men look like one another and behave in similar ways. And when we first discover some who don't—men who don't desire women, for example—we can be taken aback. Our next step is usually not to abandon the thought that men desire women, but to note an exception, while sticking to the old generalization. Only later are we likely to adopt a new category, gay men, that allows us to return to the old generalization, now about a new group, straight men. (So our second step is likely to be presupposing that everyone is either gay or straight, which turns out not to be exactly true either.)

In the course of this book, we'll encounter this most basic of our cognitive habits over and over again. So it's worth insisting from the start that essentialism about identities is usually wrong: in general, there isn't some inner essence that explains why people of a certain social identity are the way they are. We've seen already that there's more than one way to come to be a man or a woman. The story of why Ewes speak Ewe or Jains practice their religion doesn't begin with a shared inner something that explains why they do those things. And most of the things that most people do aren't done *because* they are women or men, of this or that ethnicity or race or religion. Like the imaginary Zarpies, most groups of real people, defined by the large-scale identities that shape our social world, are enormously diverse.

The Four-Day-Old Tribe

The last lesson in the psychology of identity I want to mention was demonstrated in an experiment that took place over a few days in the beautiful, hilly woods of the San Bois Mountains of Oklahoma, in 1953. That summer, a team of researchers assembled two groups of eleven-year-old boys at adjoining but separate campsites, in a place called Robbers Cave State Park. The boys were from the Oklahoma City area. They hadn't met before, but they came from similar backgrounds—they were Protestant, white, and middle-class. All this was by careful design. The researchers were studying the formation of what social psychologists call in-groups and out-groups—the way that tension developed between them and the way it might be alleviated—and the Robbers Cave experiment is a classic piece of social science.

The camp area was remote and densely wooded; the boys had been there for about a week before they learned that there was another camp of boys nearby. The two groups then challenged each other to competitive games, like baseball and tug-of-war. In the next four days, a couple of things happened. The groups gave themselves names—they were the Rattlers and the Eagles—and a fierce antagonism arose between them. Flags were torched; cabins were raided; rocks were collected as weapons for an anticipated attack.[25]

Notice that the boys felt no need for a collective name until they learned about the presence of those other boys on the campgrounds. But, as our theory predicts, to form identities they needed labels. Among the Rattlers, an ethos of 'toughness' developed, after they discovered one of the higher-status boys in the group had incurred a minor injury without mentioning it to anyone; being toughs, they also started to curse. The Eagles, having defeated the foul-mouthed Rattlers in a baseball game, decided to distinguish themselves by *not* cursing. These quasi-cultural differences could be recognized in the way each group talked about itself and the other group: the scrappy, macho Rattlers regarded the Eagles as 'sissies' and 'little babies'; the pious and clean-living Eagles considered the Rattlers to be 'bums.'[26] Labels came first, then, but essences followed fast. The boys didn't develop opposing identities because they had different norms; they developed different norms because they had opposing identities. As far as identity goes, it turns out a lot can happen in four days.

Our third psychological truth, then, is just that we humans ascribe a great deal of significance to the distinction between those who share our identities and those who don't, the insiders and the outsiders, and that we do this with identities new (like Rattlers or Eagles) and long-established, large and small, superficial and profound.

There's a whole list of psychological tendencies that go with this distinction between in-groups and out-groups. It may seem obvious, for example, that people tend to favor those of their own identity and to look down on out-group members. But given the scale of many groups, this should be more surprising than it is. Why would a Hindu give preference to another Hindu he does not know over a Muslim neighbor? There are a billion Hindus, and you have only a few hundred neighbors. And yet, everywhere in the world, we take this sort of partiality for granted.

There's a commonsense way of talking about all this. We're *clannish* creatures. We don't just belong to human kinds; we prefer our own kind and we're easily persuaded to take against outsiders. Evolutionary psychologists think these tendencies were once adaptive; they helped people survive by creating groups they could rely on to deal with the hazards of prehistoric life, including the existence of other groups competing for resources. Something like that is probably right. But whatever the explanation, it seems pretty clear that we're not just prone to essentialism, we also have these clannish tendencies, and each of us has a habitus shaped by our various identities.

The little theory of identity I just sketched and those three psychological observations helped me as I set out to think about the particular forms of identity that are the main subjects of this book. Having these ideas at hand will help us chart our way through religion, nation, race, class, and culture as sources of identity. I'm going to start with religion, because many modern religious identities connect us with some of the oldest human stories. You could debate whether, in that sense, religious identities are older than national, racial, and cultural ones; what's certain is that all of these modern forms of identity connect with religion.

In the chapters that follow, I'll be exploring a variety of ways in which identities can go awry, and can be enlisted for ill. So let me offer this stipulation as we set out: however much identity bedevils us, we cannot do without it. You'll recall the old joke. A man goes to see a psychiatrist. He says, 'Doctor, my brother's crazy—he thinks he's a chicken.' The psychiatrist says, 'Well, why don't you bring him in?' And the fellow replies, 'Oh, I would, but we need him out there laying the eggs.' Social identities may be founded in error, but they give us contours, comity, values, a sense of purpose and meaning: we need those eggs.

NOTES

1. When I quote literary classics in the text, as in the epigraphs, I won't put a reference in a note: in the age of the Internet, you'll be able to find these passages and their contexts without publishing details. When I refer to the Christian Bible, I'll use the King James translation and usually give the book. For the Quran, I'll use The *Holy Qur'an* (Classics of World Literature), trans. Abdullah Yusuf Ali (London: Wadsworth Editions, 2001). Where I have used translations, I have acknowledged the translator in the notes; where there is no translator mentioned, the translation is mine.
2. Erik H. Erikson, "Autobiographic Notes on the Identity Crisis," *Daedalus* 99 no. 4 (Fall 1970): 743, 747; Sue Erikson Bloland, *In the Shadow of Fame* (New York: Viking, 2005), 65.
3. Erik H. Erikson, *Childhood and Society*, 2nd ed. (New York: W. W. Norton, 1985; originally published 1950), 282.
4. Alvin W. Gouldner, "Cosmopolitans and Locals: Toward an Analysis of Latent Social Roles—I," *Administrative Science Quarterly* 2, no.3 (December 1957): 282–283.
5. Paul Dundas, *The Jains*, 2nd ed. (New York: Routledge & Kegan Paul, 2002), 158–159.
6. Kwame Anthony Appiah, *In My Father's House: Africa in the Philosophy of Culture* (New York: Oxford University Press, 1992).
7. In North India, the term *kinnar* does much the same work; in Tamil-Nadu, in the south, they'll often use *aravani*. *Report of the Expert Committee on the Issues Relating to Transgender Persons* (Ministry of Social Justice and Empowerment, Government of India), Appendix 2, http://socialjustice.nic.in/writereaddata/UploadFile/Binder2.pdf.
8. *Ibid.*, 102.
9. *Hijras* are not typically intersexes, it should be pointed out, although some intersexes were likely to identify as *hijras*, at least in the past.
10. Arundhati Roy, *The Ministry of Utmost Happiness* (New York: Knopf, 2017), 23.
11. Carolyn Apple, "A Navajo Worldview and Nádleehí: Implications for Western Categories," in *Two-Spirit People: Native American Gender Identity, Sexuality, and Spirituality*, ed. Sue-Ellen Jacobs, Wesley Thomas, and Sabine Lang (Champaign: University of Illinois Press, 1997), 174–191.
12. M. Blackless et al., "How Sexually Dimorphic Are We? Review and Synthesis," *American Journal of Human Biology* 12, no. 2 (March 2000): 151–166. And cf. "How Common Is Intersex," Intersex Society of North America, http://www.isna.org/faq/frequency. Scientific estimates depend on which conditions count as intersexual, and range widely, from .05 to 1.7 percent.

13. Kimberlé W. Crenshaw, "Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color," *Stanford Law Review* 43 no. 6 (July 1991): 1241–1299.
14. See, e.g., "The Habitus and the Space of Life-Styles," in Pierre Bourdieu, *Distinctions* (Cambridge: Harvard University Press, 1987),169 et seq.; and John B. Thompson, "Editor's Introduction," in Pierre Bourdieu, *Language and Symbolic Power* (Cambridge: Polity Press, 1991), 12.
15. Pierre Bourdieu, *The Logic of Practice* (Stanford: Stanford University Press, 1980), 69–70; Thompson, in Bourdieu, *Language and Symbolic Power*, 13.
16. *". . . un nouveau parler des intellectuels, un peu hésitant, voire bredouillant, interrogatif ('non?') et entrecoupé"; ". . . l'ancien usage professoral (avec ses périodes, ses imparfaits du subjonctif, etc.)," Pierre Bourdieu, Ce que parler veut dire: L'économie des échanges linguistiques* (Paris: Fayard, 1982), 56.
17. *The Encyclopedia of African-American Popular Culture*, ed. Jesse Carney Smith (Santa Barbara, CA: Greenwood, 2011), 3:1089.
18. Bourdieu, *Language and Symbolic Power*, 86–89.
19. Thompson, in Bourdieu, *Language and Symbolic Power*, 17–18.
20. Claude Steele, *Whistling Vivaldi: And Other Clues to How Stereotypes Affect Us* (New York: W. W. Norton, 2010), 7. The graduate student in question, Brent Staples, went on to be a well-known *New York Times* columnist.
21. Susan Gelman, "Psychological Essentialism in Children," *Trends in Cognitive Sciences* 8 no. 9 (September 2004): 404–409.
22. Sarah-Jane Leslie, "The Original Sin of Cognition: Fear, Prejudice, and Generalization," *Journal of Philosophy* 114 no. 8 (August 2017): 393–421.
23. Marjorie Rhodes, Sarah-Jane Leslie, and Christina M. Tworek, "Cultural Transmission of Social Essentialism," *Proceedings of the National Academy of Sciences* 109 no. 34 (August 21, 2012): 13526–13531. The idea that some racial prejudice arises from the normal operation of basic cognitive mechanisms is the thrust of Sarah-Jane Leslie's argument in "The Original Sin."
24. Graham Robb, *The Discovery of France: A Historical Geography* (New York: W. W. Norton, 2007), 43–47. Robb sorts through various old conjectures—were the Cagots descendants of Visigoths, or Saracen invaders, or Cathar heretics, or lepers?—and theorizes that they may have originated simply as a medieval carpenter's guild, stigmatized by rival guilds. In the nineteenth century, the widely traveled Irish writer Thomas Colley Grattan described them as "a rejected caste, enveloped in a mystery which no research can penetrate," and proposed, "We might, by imagining the possibility of their amelioration, lead to plans for their relief; and instead of useless efforts to account for their miseries, make practical attempts to remove them. . . . While contemplating with shame, the narrow circle into which man may imprison his fellow man, we have at least the consolation of knowing that he possesses in himself the power of dissolving the shackles he has forged; and of buying the memory of his own injustice, in an oblivious flood of charity and atonement." Thomas Colley Grattan, *High-ways and By-ways: The Cagot's Hut* (London: Henry Colburn & Richard Bentley, 1831), 46, 48–49. See also Sean Thomas, "The Last Untouchable in Europe," *The Independent on Sunday*, July 27, 2008, http://www.independent.co.uk/news/world/europe/the-last-untouchable-in-europe-878705.html.
25. Muzafer Sherif et al., *The Robbers Cave Experiment: Intergroup Conflict and Cooperation* (Middletown, CT: Wesleyan University Press, 1988; originally published by the Institute of Group Relations, University of Oklahoma, 1961), 95–116. I discussed this experiment in my book *The Ethics of Identity* (Princeton: Princeton University Press, 2005), 84–85.
26. Sherif et al., *op. cit.*, 116.

7 Artificial Intelligence

Can there be artificial minds? There is a thriving industry of artificial intelligence, devoted to constructing machines that can perform tasks that would require intelligence when performed by humans. AI systems have gone from strength to strength in domains such as image recognition, speech recognition, language translation, game-playing, and much more. The question then arises: do any of these AI systems have a *mind*? Do they genuinely *perceive* or *think* or *understand*? If they do not, might an AI system one day?

Alan Turing (chapter 75) is often credited with founding the field of artificial intelligence with his 1950 article on "Computing Machinery and Intelligence." He said the question of 'Can machines think?' is too meaningless to deserve further discussion, and he proposed to replace it with another. Turing's question focuses on a test that he called the Imitation Game and that has come to be called the Turing test. To pass the Turing test, a machine needs to be able to carry on a humanlike conversation that a human judge cannot distinguish from a real human conversation with greater than 50 percent accuracy. Turing suggested that machines could eventually pass this test, and that if they did they could reasonably be said to have minds.

John Searle (chapter 76) argues that while some machines might have minds (we may ourselves be machines), these minds cannot be grounded in computation alone: merely programming a computer in the right way will not suffice for a mind. He argues for this using a thought-experiment about a room in which an English speaker simulates a computer program for understanding Chinese, without any real understanding. Searle argues that computers are in the same situation: they have syntax, but no real semantics. The conclusion is that running a computer program does not suffice to have genuine mental states. Searle holds that to have a mind, a system must have the causal powers of the human brain.

A common version of the anti-AI view holds that only a biological system can have a mind, and that a silicon machine could not in principle. This view is parodied in Terry Bisson's story "They're Made Out of Meat?" (chapter 74). A group of machines discover the earth, and are horrified to see flesh-based intelligence. They have a hard time seriously entertaining the idea that there can be thinking meat. From our perspective they seem biased and prejudiced. This raises the question of whether human doubts about machine consciousness are biased in a similar way.

The last three chapters discuss the future of artificial intelligence. My contribution in chapter 77 focuses on the 'singularity': a scenario in which machines become more intelligent than humans, and then create even more intelligent machines, with a rapid spiral to superintelligence. This idea is often dismissed as science fiction, but I suggest

that it is supported by a strong philosophical argument. If certain conditions are satisfied, a singularity is very likely. As a result, there are only a limited number of ways in which a singularity can be avoided.

Nick Bostrom and Eliezer Yudkowsky (chapter 78) discuss ethical issues about artificial intelligence. Some ethical issues apply to current AI systems: whether we can avoid unfair bias in these systems, whether we can understand how they work, and whether we can avoid being manipulated by them. Other issues apply to future systems with artificial general intelligence. One key issue is whether we can design these systems in a way that is safe and beneficial for the human population. Another issue is whether those systems may have moral status themselves. A final question is whether we could or should someday upload our minds so that we are transferred to a digital computer.

The final chapter by Susan Schneider and Pete Mandik discusses ways in which the philosophy of mind may be highly relevant in our technological future. They discuss scenarios including brain enhancement (raising issues about the extended mind), mind uploading (raising issues about personal identity), and artificial consciousness (raising issues about the problem of consciousness). In each case, philosophical reasoning may be essential for dealing with these technologies in an optimal way.

FURTHER READING

Shieber 2004 collects many important articles on the Turing test. Hofstadter 1977 is a classic book on artificial and human minds. Lucas 1961 and Penrose 1994 give well-known arguments for limitations on artificial intelligence. Bostrom 1992, Searle 1992, and Schneider 2019 give book-length treatments of their views. Awret 2016 collects twenty-six responses to my article on the singularity along with my reply. Liao (2020) is a collection addressing many issues in the ethics of artificial intelligence.

Awret, U., ed., *The Singularity* (Exeter, UK: Imprint Academic, 2016).

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (Oxford, UK: Oxford University Press, 2014).

Hofstadter, D. R. *Godel, Escher, Bach: An Eternal Golden Braid* (New York: Basic Books, 1977).

Liao, M. *The Ethics of Artificial Intelligence* (New York: Oxford University Press, 2020).

Lucas, J. R., "Minds, machines, and Gödel," *Philosophy* 36 (1961): pp. 112–37.

Penrose, R. *Shadows of the Mind* (Oxford, UK: Oxford University Press, 1994).

Searle, J. *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992).

Shieber, S. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence* (Cambridge, MA: MIT Press, 2004).

Schneider, S. *Artificial You* (Princeton, NJ: Princeton University Press, 2019).

They're Made Out of Meat

Terry Bisson

'They're made out of meat.'

'Meat?'

'Meat. They're made out of meat.'

'Meat?'

'There's no doubt about it. We picked up several from different parts of the planet, took them aboard our recon vessels, and probed them all the way through. They're completely meat.'

'That's impossible. What about the radio signals? The messages to the stars?'

'They use the radio waves to talk, but the signals don't come from them. The signals come from machines.'

'So who made the machines? That's who we want to contact.'

'They made the machines. That's what I'm trying to tell you. Meat made the machines.'

'That's ridiculous. How can meat make a machine? You're asking me to believe in sentient meat.'

'I'm not asking you, I'm telling you. These creatures are the only sentient race in that sector and they're made out of meat.'

'Maybe they're like the orfolei. You know, a carbon-based intelligence that goes through a meat stage.'

'Nope. They're born meat and they die meat. We studied them for several of their life spans, which didn't take long. Do you have any idea what's the life span of meat?'

'Spare me. Okay, maybe they're only part meat. You know, like the weddilei. A meat head with an electron plasma brain inside.'

'Nope. We thought of that, since they do have meat heads, like the weddilei. But I told you, we probed them. They're meat all the way through.'

'No brain?'

'Oh, there's a brain all right. It's just that the brain is made out of meat! That's what I've been trying to tell you.'

'So . . . what does the thinking?'

'You're not understanding, are you? You're refusing to deal with what I'm telling you. The brain does the thinking. The meat.'

'Thinking meat! You're asking me to believe in thinking meat!'

'Yes, thinking meat! Conscious meat! Loving meat. Dreaming meat. The meat is the whole deal! Are you beginning to get the picture or do I have to start all over?'

'Omigod. You're serious then. They're made out of meat.'

'Thank you. Finally. Yes. They are indeed made out of meat. And they've been trying to get in touch with us for almost a hundred of their years.'

'Omigod. So what does this meat have in mind?'

'First it wants to talk to us. Then I imagine it wants to explore the Universe, contact other sentiences, swap ideas and information. The usual.'

'We're supposed to talk to meat.'

'That's the idea. That's the message they're sending out by radio. 'Hello. Anyone out there. Anybody home.' That sort of thing.'

'They actually do talk, then. They use words, ideas, concepts?' 'Oh, yes. Except they do it with meat.'

'I thought you just told me they used radio.'

'They do, but what do you think is on the radio? Meat sounds. You know how when you slap or flap meat, it makes a noise? They talk by flapping their meat at each other. They can even sing by squirting air through their meat.'

'Omigod. Singing meat. This is altogether too much. So what do you advise?'

'Officially or unofficially?'

'Both.'

'Officially, we are required to contact, welcome and log in any and all sentient races or multibeings in this quadrant of the Universe, without prejudice, fear or favor. Unofficially, I advise that we erase the records and forget the whole thing.'

'I was hoping you would say that.'

'It seems harsh, but there is a limit. Do we really want to make contact with meat?'

'I agree one hundred percent. What's there to say? 'Hello, meat. How's it going?' But will this work? How many planets are we dealing with here?'

'Just one. They can travel to other planets in special meat containers, but they can't live

From *Omni*, April 1991. Reprinted with permission of the author.

on them. And being meat, they can only travel through C space. Which limits them to the speed of light and makes the possibility of their ever making contact pretty slim. Infinitesimal, in fact.'

'So we just pretend there's no one home in the Universe.'

'That's it.'

'Cruel. But you said it yourself, who wants to meet meat? And the ones who have been aboard our vessels, the ones you probed? You're sure they won't remember?'

'They'll be considered crackpots if they do. We went into their heads and smoothed out their meat so that we're just a dream to them.'

'A dream to meat! How strangely appropriate, that we should be meat's dream.'

'And we marked the entire sector unoccupied.'

'Good. Agreed, officially and unofficially. Case closed. Any others? Anyone interesting on that side of the galaxy?'

'Yes, a rather shy but sweet hydrogen core cluster intelligence in a class nine star in G445 zone. Was in contact two galactic rotations ago, wants to be friendly again.'

'They always come around.'

'And why not? Imagine how unbearably, how unutterably cold the Universe would be if one were all alone. . .'

Computing Machinery and Intelligence

A. M. Turing

1. The Imitation Game

I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think.' The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult, to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game.' It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A.' The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be

> 'My hair is shingled, and the longest strands are about nine inches long.'

In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as 'I am the woman, don't listen to him!' to her

From *Mind* 59: 433–60, 1950. Reprinted by permission of the publisher.

answers, but it will avail nothing as the man can make similar remarks.

We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'

2. Critique of the New Problem

As well as asking, 'What is the answer to this new form of the question' one may ask, 'Is this new question a worthy one to investigate?' This latter question we investigate without further ado, thereby cutting short an infinite regress.

The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man. No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin. It is possible that at some time this might be done, but even supposing this invention available we should feel there was little point in trying to make a 'thinking machine' more human by dressing it up in such artificial flesh. The form in which we have set the problem reflects this fact in the condition which prevents the interrogator from seeing or touching the other competitors, or hearing their voices. Some other advantages of the proposed criterion may be shown up by specimen questions and answers. Thus:

Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764
A: (Pause about 30 seconds and then give as answer) 105621.
Q: Do you play chess?
A: Yes.
Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds) R-R8 mate.

The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include. We do not wish to penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against an aeroplane. The conditions of our game make these disabilities irrelevant. The 'witnesses' can brag, if they consider it advisable, as much as they please about their charms, strength or heroism, but the interrogator cannot demand practical demonstrations.

The game may perhaps be criticised on the ground that the odds are weighted too heavily against the machine. If the man were to try and pretend to be the machine he would clearly make a very poor showing. He would be given away at once by slowness and inaccuracy in arithmetic. May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection.

It might be urged that when playing the 'imitation game' the best strategy for the machine may possibly be something other than imitation of the behaviour of a man. This may be, but I think it is unlikely that there is any great effect of this kind. In any case there is no intention to investigate here the theory of the game, and it will be assumed that the best strategy is to try to provide answers that would naturally be given by a man.

3. The Machines Concerned in the Game

The question which we put in section 1 will not be quite definite until we have specified what we mean by the word 'machine.' It is natural that we should wish to permit every kind of engineering technique to be used in our machines. We also wish to allow the possibility than an engineer or team of engineers may construct a machine which works, but whose manner of operation cannot be satisfactorily described by its constructors because they have applied a method which is largely experimental. Finally, we wish to exclude from the machines men born in the usual manner. It is difficult to frame the definitions so as to satisfy these three conditions. One might for instance insist that the team of engineers should be all of one sex, but this would not really be satisfactory, for it is probably possible to rear a complete individual from

a single cell of the skin (say) of a man. To do so would be a feat of biological technique deserving of the very highest praise, but we would not be inclined to regard it as a case of constructing a 'thinking machine.' This prompts us to abandon the requirement that every kind of technique should be permitted. We are the more ready to do so in view of the fact that the present interest in 'thinking machines' has been aroused by a particular kind of machine, usually called an 'electronic computer' or 'digital computer.' Following this suggestion we only permit digital computers to take part in our game.

This restriction appears at first sight to be a very drastic one. I shall attempt to show that it is not so in reality. To do this necessitates a short account of the nature and properties of these computers.

It may also be said that this identification of machines with digital computers, like our criterion for 'thinking' will only be unsatisfactory if (contrary to my belief), it turns out that digital computers are unable to give a good showing in the game.

There are already a number of digital computers in working order, and it may be asked, 'Why not try the experiment straight away? It would be easy to satisfy the conditions of the game. A number of interrogators could be used, and statistics compiled to show how often the right identification was given.' The short answer is that we are not asking whether all digital computers would do well in the game nor whether the computers at present available would do well, but whether there are imaginable computers which would do well. But this is only the short answer. We shall see this question in a different light later.

4. Digital Computers

The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer. The human computer is supposed to be following fixed rules; he has no authority to deviate from them in any detail. We may suppose that these rules are supplied in a book, which is altered whenever he is put on to a new job. He has also an unlimited supply of paper on which he does his calculations. He may also do his multiplications and additions on a 'desk machine' but this is not important.

If we use the above explanation as a definition we shall be in danger of circularity of argument. We avoid this by giving an outline of the means by which the desired effect is achieved. A digital computer can usually be regarded as consisting of three parts:

(i) Store.
(ii) Executive unit.
(iii) Control.

The store is a store of information, and corresponds to the human computer's paper, whether this is the paper on which he does his calculations or that on which his book of rules is printed. In so far as the human computer does calculations in his head a part of the store will correspond to his memory.

The executive unit is the part which carries out the various individual operations involved in a calculation. What these individual operations are will vary from machine to machine. Usually fairly lengthy operations can be done such as 'Multiply 3540675445 by 7076345687' but in some machines only very simple ones such as 'Write down 0' are possible.

We have mentioned that the 'book of rules' supplied to the computer is replaced in the machine by a part of the store. It is then called the 'table of instructions.' It is the duty of the control to see that these instructions are obeyed correctly and in the right order. The control is so constructed that this necessarily happens.

The information in the store is usually broken up into packets of moderately small size. In one machine, for instance, a packet might consist of ten decimal digits. Numbers are assigned to the parts of the store in which the various packets of information are stored, in some systematic manner. A typical instruction might say—

'Add the number stored in position 6809 to that in 4302 and put the result back into the latter storage position.'

Needless to say it would not occur in the machine expressed in English. It would more likely be coded in a form such as 6809430217. Here 17 says which of various possible operations is to be performed on the two numbers. In this case the operation is that described above, *viz.* 'Add the number . . .' It will be noticed that the instruction takes up 10 digits and so forms one packet of information, very conveniently. The control will normally take the instructions to be obeyed in the order of the positions in which they are stored, but occasionally an instruction such as

'Now obey the instruction stored in position 5606, and continue from there' may be encountered, or again

'If position 4505 contains 0 obey next the instruction stored in 6707, otherwise continue straight on.' Instructions of these latter types are very important because they make it possible for a sequence of operations to be repeated over and over again until some condition is fulfilled, but in doing so to obey, not fresh instructions on each repetition, but the same ones over and over again. To take a domestic analogy. Suppose Mother wants Tommy to call at the cobbler's every morning on his way to school to see if her shoes are done, she can ask him afresh every morning. Alternatively she can stick up a notice once and for all in the hall which he will see when he leaves for school and which tells him to call for the shoes, and also to destroy the notice when he comes back if he has the shoes with him.

The reader must accept it as a fact that digital computers can be constructed, and indeed have been constructed, according to the principles we have described, and that they can in fact mimic the actions of a human computer very closely.

The book of rules which we have described our human computer as using is of course a convenient fiction. Actual human computers really remember what they have got to do. If one wants to make a machine mimic the behaviour of the human computer in some complex operation one has to ask him how it is done, and then translate the answer into the form of an instruction table. Constructing instruction tables is usually described as 'programming.' To 'programme a machine to carry out the operation A' means to put the appropriate instruction table into the machine so that it will do A.

An interesting variant on the idea of a digital computer is a 'digital computer with a random element.' These have instructions involving the throwing of a die or some equivalent electronic process; one such instruction might for instance be, 'Throw the die and put the resulting number into store 1000.' Sometimes such a machine is described as having free will (though I would not use this phrase myself). It is not normally possible to determine from observing a machine whether it has a random element, for a similar effect can be produced by such devices as making the choices depend on the digits of the decimal for π.

Most actual digital computers have only a finite store. There is no theoretical difficulty in the idea of a computer with an unlimited store. Of course only a finite part can have been used at any one time. Likewise only a finite amount can have been constructed, but we can more and more being added as required. Such computers have special theoretical interest and will be called infinitive capacity computers.

The idea of a digital computer is an old one. Charles Babbage, Lucasian Professor of Mathematics at Cambridge from 1828 to 1839, planned such a machine, called the Analytical Engine, but it was never completed. Although Babbage had all the essential ideas, his machine was not at that time such a very attractive prospect. The speed which would have been available would be definitely faster than a human computer but something like 100 times slower than the Manchester machine, itself one of the slower of the modern machines. The storage was to be purely mechanical, using wheels and cards.

The fact that Babbage's Analytical Engine was to be entirely mechanical will help us to rid ourselves of a superstition. Importance is often attached to the fact that modern digital computers are electrical, and that the nervous system also is electrical. Since Babbage's machine was not electrical, and since all digital computers are in a sense equivalent, we see that this use of electricity cannot be of theoretical importance. Of course electricity usually comes in where fast signaling is concerned, so that it is not surprising that we find it in both these connections. In the nervous system chemical phenomena are at least as important as electrical. In certain computers the storage system is mainly acoustic. The feature of using electricity is thus seen to be only a very superficial similarity. If we wish to find such similarities we should look rather for mathematical analogies of function.

5. Universality of Digital Computers

The digital computers considered in the last section may be classified amongst the 'discrete state machines.' These are the machines which move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. Strictly speaking there are no such machines. Everything really moves continuously. But there are many kinds of machine which can profitably be *thought of* as being discrete state machines. For instance in considering the switches for a lighting system it

is a convenient fiction that each switch must be definitely on or definitely off. There must be intermediate positions, but for most purposes we can forget about them. As an example of a discrete state machine we might consider a wheel which clicks round through 120° once a second, but may be stopped by a lever which can be operated from outside; in addition a lamp is to light in one of the positions of the wheel. This machine could be described abstractly as follows. The internal state of the machine (which is described by the position of the wheel) may be q_1, q_2 or q_3. There is an input signal i_0 or i_1 (position of lever). The internal state at any moment is determined by the last state and input signal according to the table

Input	q_1	q_2	q_3
i_0	q_2	q_3	q_1
i_1	q_1	q_2	q_3

The output signals, the only externally visible indication of the internal state (the light) are described by the table

State	q_1	q_2	q_3
Output	o_0	o_0	o_1

This example is typical of discrete state machines. They can be described by such tables provided they have only a finite number of possible states.

It will seem that given the initial state of the machine and the input signals it is always possible to predict all future states. This is reminiscent of Laplace's view that from the complete state of the universe at one moment of time, as described by the positions and velocities of all particles, it should be possible to predict all future states. The prediction which we are considering is, however, rather nearer to practicability than that considered by Laplace. The system of the 'universe as a whole' is such that quite small errors in the initial conditions can have an overwhelming effect at a later time. The displacement of a single electron by a billionth of a centimetre at one moment might make the difference between a man being killed by on avalanche a year later, or escaping. It is an essential property of the mechanical systems which we have called 'discrete state machines' that this phenomenon does not occur. Even when we consider the actual physical machines instead of the idealised machines, reasonably accurate knowledge of the state at one moment yields reasonably accurate knowledge any number of steps later.

As we have mentioned, digital computers fall within the class of discrete state machines. But the number of states of which such a machine is capable is usually enormously large. For instance, the number for the machine now working at Manchester is about $2^{165,000}$, *i.e.* about $10^{50,000}$. Compare this with our example of the clicking wheel described above, which had three states. It is not difficult to see why the number of states should be so immense. The computer includes a store corresponding to the paper used by a human computer. It must be possible to write into the store any one of the combinations of symbols which might have been written on the paper. For simplicity suppose that only digits from 0 to 9 are used as symbols. Variations in handwriting are ignored. Suppose the computer is allowed 100 sheets of paper each containing 60 lines each with room for 30 digits. Then the number of states is $10^{100\times50\times30}$, *i.e.* $10^{150,000}$. This is about the number of states of three Manchester machines put together. The logarithm to the base two of the number of states is usually called the 'storage capacity' of the machine. Thus the Manchester machine has a storage capacity of about 165,000 and the wheel machine of our example about 1·6. If two machines are put together their capacities must be added to obtain the capacity of the resultant machine. This leads to the possibility of statements such as 'The Manchester machine contains 64 magnetic tracks each with a capacity of 2560, eight electronic tubes with a capacity of 1,280. Miscellaneous storage amounts to about 300 making a total of 174,380.'

Given the table corresponding to a discrete state machine it is possible to predict what it will do. There is no reason why this calculation should not be carried out by means of a digital computer. Provided it could be carried out sufficiently quickly the digital computer could mimic the behaviour of any discrete state machine. The imitation game could then be played with the machine in question (as B) and the mimicking digital computer (as A) and the interrogator would be unable to distinguish them. Of course the digital computer must have an adequate storage capacity as well as working sufficiently fast. Moreover, it must be programmed afresh for each new machine which it is desired to mimic.

This special property of digital computers, that they can mimic any discrete state machine, is described by saying that they are *universal*

machines. The existence of machines with this property has the important consequence that, considerations of speed apart, it is unnecessary to design various new machines to do various computing processes. They can all be done with one digital computer, suitably programmed for each case. It will be seen that as a consequence of this all digital computers are in a sense equivalent.

We may now consider again the point raised at the end of section 3. It was suggested tentatively that the question, 'Can machines think?' should be replaced by 'Are there imaginable digital computers which would do well in the imitation game?' If we wish we can make this superficially more general and ask 'Are there discrete state machines which would do well?' But in view of the universality property we see that either of these questions is equivalent to this, 'Let us fix our attention on one particular digital computer *C*. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, *C* can be made to day satisfactorily the part of A in the imitation game, the part of B being taken by a man?'

6. Contrary Views on the Main Question

We may now consider the ground to have been cleared and we are ready to proceed to the debate on our question, 'Can machines think?' and the variant of it quoted at the end of the last section. We cannot altogether abandon the original form of the problem, for opinions will differ as to the appropriateness of the substitution and we must at least listen to what has to be said in this connexion.

It will simplify matters for the reader if I explain first my own beliefs in the matter. Consider first the more accurate form of the question. I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning. The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. I believe further that no useful purpose is served by concealing these beliefs. The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any unproved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful linen of research.

I now proceed to consider opinions opposed to my own.

(1) *The Theological Objection.* Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think.[1]

I am unable to accept any part of this, but will attempt to reply in theological terms. I should find the argument more convincing if animals were classed with men, for there is a greater difference, to my mind, between the typical animate and the inanimate than there is between man and the other animals. The arbitrary character of the orthodox view becomes clearer if we consider how it might appear to a member of some other religious community. How do Christians regard the Moslem view that women have no souls? But let us leave this point aside and return to the main argument. It appears to me that the argument quoted above implies a serious restriction of the omnipotence of the Almighty. It is admitted that there are certain things that He cannot do such as making one equal to two, but should we not believe that He has freedom to confer a soul on an elephant if He sees fit? We might expect that He would only exercise this power in conjunction with a mutation which provided the elephant with an appropriately improved brain to minister to the needs of this soul. An argument of exactly similar form may be made for the case of machines. It may seem different because it is more difficult to 'swallow.' But this really only means that we think it would be less likely that He would consider the circumstances suitable for conferring a soul. The circumstances in question are discussed in the rest of this paper. In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates.

However, this is mere speculation. I am not very impressed with theological arguments whatever they may be used to support. Such arguments have often been found unsatisfactory in the past. In the time of Galileo it was argued that the texts, 'And the sun stood still . . . and hasted not to go down about a whole day' (Joshua x. 13) and 'He laid the foundations of the earth, that it should not move at any time' (Psalm cv. 5) were an adequate refutation of the Copernican theory. With our present knowledge such an argument appears futile. When that knowledge was not available it made a quite different impression.

(2) *The 'Heads in the Sand' Objection.* 'The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so.'

This argument is seldom expressed quite so openly as in the form above. But it affects most of us who think about it at all. We like to believe that Man is in some subtle way superior to the rest of creation. It is best if he can be shown to be *necessarily* superior, for then there is no danger of him losing his commanding position. The popularity of the theological argument is clearly connected with this feeling. It is likely to be quite strong in intellectual people, since they value the power of thinking more highly than others, and are more inclined to base their belief in the superiority of Man on this power.

I do not think that this argument is sufficiently substantial to require refutation. Consolation would be more appropriate: perhaps this should be sought in the transmigration of souls.

(3) *The Mathematical Objection.* There are a number of results of mathematical logic which can be used to show that there are limitations to the powers of discrete-state machines. The best known of these results is known as Gödel's theorem,[2] and shows that in any sufficiently powerful logical system statements can be formulated which can neither be proved nor disproved within the system, unless possibly the system itself is inconsistent. There are other, in some respects similar, results due to *Church*, *Kleene*, *Rosser*, and *Turing*. The latter result is the most convenient to consider, since it refers directly to machines, whereas the others can only be used in a comparatively indirect argument: for instance if Gödel's theorem is to be used we need in addition to have some means of describing logical systems in terms of machines, and machines in terms of logical systems. The result in question refers to a type of machine which is essentially a digital computer with an infinite capacity. It states that there are certain things that such a machine cannot do. If it is rigged up to give answers to questions as in the imitation game, there will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply. There may, of course, be many such questions, and questions which cannot be answered by one machine may be satisfactorily answered by another. We are of course supposing for the present that the questions are of the kind to which an answer 'Yes' or 'No' is appropriate, rather than questions such as 'What do you think of Picasso?' The questions that we know the machines must fail on are of this type, 'Consider the machine specified as follows . . . Will this machine ever answer 'Yes' to any question?" The dots are to be replaced by a description of some machine in a standard form, which could be something like that used in section 5. When the machine described bears a certain comparatively simple relation to the machine which is under interrogation, it can be shown that the answer is either wrong or not forthcoming. This is the mathematical result: it is argued that it proves a disability of machines to which the human intellect is not subject.

The short answer to this argument is that although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect. But I do not think this view can be dismissed quite so lightly. Whenever one of these machines is asked the appropriate critical question, and gives a definite answer, we know that this answer must be wrong, and this gives us a certain feeling of superiority. Is this feeling illusory? It is no doubt quite genuine, but I do not think too much importance should be attached to it. We too often give wrong answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines. Further, our superiority can only be felt on such an occasion in relation to the one machine over which we have scored our petty triumph. There would be no question of triumphing simultaneously over *all* machines. In short, then, there might be men cleverer than any given machine, but then again there might be other machines cleverer again, and so on.

Those who hold to the mathematical argument would, I think, mostly be willing to accept the imitation game as a basis for discussion. Those who believe in the two previous

objections would probably not be interested in any criteria.

(4) *The Argument from Consciousness.* This argument is very well expressed in *Professor Jefferson*'s Lister Oration for 1949, from which I quote. 'Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants.'

This argument appears to be a denial of the validity of our test. According to the most extreme form of this view the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise according to this view the only way to know that a *man* thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe 'A thinks but B does not' whilst B believes 'B thinks but A does not.' Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks.

I am sure that Professor Jefferson does not wish to adopt the extreme and solipsist point of view. Probably he would be quite willing to accept the imitation game as a test. The game (with the player B omitted) is frequently used in practice under the name of *viva voce* to discover whether some one really understands something or has 'learnt it parrot fashion.' Let us listen in to a part of such a *viva voce*:

Interrogator: In the first line of your sonnet which reads 'Shall I compare thee to a summer's day would not 'a spring day' do as well or better?

Witness: It wouldn't scan.

Interrogator: How about 'a winter's day'? That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

And so on. What would Professor Jefferson say if the sonnet-writing machine was able to answer like this in the *viva voce?* I do not know whether he would regard the machine as 'merely artificially signalling' these answers, but if the answers were as satisfactory and sustained as in the above passage I do not think he would describe it as 'an easy contrivance.' This phrase is, I think, intended to cover such devices as the inclusion in the machine of a record of someone reading a sonnet, with appropriate switching to turn it on from time to time.

In short then, I think that most of those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position. They will then probably be willing to accept our test.

I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper.

(5) *Arguments from Various Disabilities.* These arguments take the form, 'I grant you that you can make machines do all the things you have mentioned but you will never be able to make one to do X.' Numerous features X are suggested in this connexion. I offer a selection:

Be kind, resourceful, beautiful, friendly (p. 448), have initiative, have a sense of humour, tell right from wrong, make mistakes (p. 448), fall in love, enjoy strawberries and cream (p. 448), make someone fall in love with it, learn from experience (pp. 466 f.), use words properly, be the subject of its own thought (p. 449), have as much diversity of behaviour as a man, do something really new (p. 460). (Some of these disabilities are given special consideration as indicated by the page numbers.)

No support is usually offered for these statements. I believe they are mostly founded on the principle of scientific induction. A man has seen thousands of machines in his lifetime. From what he sees of them he draws a number of general conclusions. They are ugly, each is

designed for a very limited purpose, when required for a minutely different purpose they are useless, the variety of behaviour of any one of them is very small, etc., etc. Naturally he concludes that these are necessary properties of machines in general. Many of these limitations are associated with the very small storage capacity of most machines. (I am assuming that the idea of storage capacity is extended in some way to cover machines other than discrete-state machines. The exact definition does not matter as no mathematical accuracy is claimed in the present discussion.) A few years ago, when very little had been heard of digital computers, it was possible to elicit much incredulity concerning them, if one mentioned their properties without describing their construction. That was presumably due to a similar application of the principle of scientific induction. These applications of the principle are of course largely unconscious. When a burnt child fears the fire and shows that he fears it by avoiding it, I should say that he was applying scientific induction. (I could of course also describe his behaviour in many other ways.) The works and customs of mankind do not seem to be very suitable material to which to apply scientific induction. A very large part of space-time must he investigated, if reliable results are to be obtained. Otherwise we may (as most English children do) decide that everybody speaks English, and that it is silly to learn French.

There are, however, special remarks to be made about many of the disabilities that have been mentioned. The inability to enjoy strawberries and cream may have struck the reader as frivolous. Possibly a machine might be made to enjoy this delicious dish, but any attempt to make one do so would be idiotic. What is important about this disability is that it contributes to some of the other disabilities, *e.g.* to the difficulty of the same kind of friendliness occurring between man and machine as between white man and white man, or between black man and black man.

The claim that 'machines cannot make mistakes' seems a curious one. One is tempted to retort, 'Are they any the worse for that?' But let us adopt a more sympathetic attitude, and try to see what is really meant. I think this criticism can be explained in terms of the imitation game. It is claimed that the interrogator could distinguish the machine from the man simply by setting them a number of problems in arithmetic. The machine would be unmasked because of its deadly accuracy. The reply to this is simple. The machine (programmed for playing the game) would not attempt to give the *right* answers to the arithmetic problems. It would deliberately introduce mistakes in a manner calculated to confuse the interrogator. A mechanical fault would probably show itself through an unsuitable decision as to what sort of a mistake to make in the arithmetic. Even this interpretation of the criticism is not sufficiently sympathetic. But we cannot afford the space to go into it much further. It seems to me that this criticism depends on a confusion between two kinds of mistake. We may call them 'errors of functioning' and 'errors of conclusion.' Errors of functioning are due to Borne mechanical or electrical fault which causes the machine to behave otherwise than it was designed to do. In philosophical discussions one likes to ignore the possibility of such errors; one is therefore discussing 'abstract machines.' These abstract machines are mathematical fictions rather than physical objects. By definition they are incapable of errors of functioning. In this sense we can truly say that 'machines can never make mistakes.' Errors of conclusion can only arise when some meaning is attached to the output signals from the machine. The machine might, for instance, type out mathematical equations, or sentences in English. When a false proposition is typed we say that the machine has committed an error of conclusion. There is clearly no reason at all for saying that a machine cannot make this kind of mistake. It might do nothing but type out repeatedly '0=1.' To take a lees perverse example, it might have some method for drawing conclusions by scientific induction. We must expect such a method to lead occasionally to erroneous results.

The claim that a machine cannot be the subject of its own thought can of course only be answered if it can be shown that the machine has *some* thought with *some* subject matter. Nevertheless, 'the subject matter of a machine's operations' does seem to mean something, at least to the people who deal with it. If, for instance, the machine was trying to find a solution of the equation $x^2 - 40x - 11 = 0$ one would be tempted to describe this equation as part of the machine's subject matter at that moment. In this sort of sense a machine undoubtedly can be its own subject matter. It may be used to help in making up its own programmes, or to predict the effect of alterations in its own structure. By observing the results of its own behaviour it can modify its own programmes so as to achieve some purpose more effectively. These

are possibilities of the near future, rather than Utopian dreams.

The criticism that a machine cannot have much diversity of behaviour is just a way of saying that it cannot have much storage capacity. Until fairly recently a storage capacity of even a thousand digits was very rare.

The criticisms that we are considering here are often disguised forms of the argument from consciousness. Usually if one maintains that a machine *can* do one of these things, and describes the kind of method that the machine could use, one will not make much of an impression. It is thought that the method (whatever it may be, for it must be mechanical) is really rather base.

(6) *Lady Lovelace's Objection.* Our most detailed information of Babbage's Analytical Engine comes from a memoir by *Lady Lovelace*. In it she states, 'The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform' (her italics). This statement is quoted by *Hartree* who adds: 'This does not imply that it may not be possible to construct electronic equipment which will 'think for itself' or in which, in biological terms, one could set up a conditioned reflex, which would serve as a basis for 'learning.' Whether this is possible in principle or not is a stimulating and exciting question, suggested by some of these recent developments. But it did not seem that the machines constructed or projected at the time had this property.'

I am in thorough agreement with Hartree over this. It will be noticed that he does not assert that the machines in question had not got the property, but rather that the evidence available to Lady Lovelace did not encourage her to believe that they had it. It is quite possible that the machines in question had in a sense got this property. For suppose that some discrete–state machine has the property. The Analytical Engine was a universal digital computer, so that, if its storage capacity and speed were adequate, it could by suitable programming be made to mimic the machine in question. Probably this argument did not occur to the Countess or to Babbage. In any case there was no obligation on them to claim all that could be claimed.

This whole question will be considered again under the heading of learning machines.

A variant of Lady Lovelace's objection states that a machine can 'never do anything really new.' This may be parried for a moment with the saw, 'There is nothing new under the sun.' Who can be certain that 'original work' that he has done was not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles. A better variant of the objection says that a machine can never 'take us by surprise.' This statement is a more direct challenge and can be met directly. Machines take me by surprise with great frequency. This is largely because I do not do sufficient calculation to decide what to expect them to do, or rather because, although I do a calculation, I do it in a hurried, slipshod fashion, taking risks. Perhaps I say to myself, 'I suppose the voltage here ought to be the same as there: anyway let's assume it is.' Naturally I am often wrong, and the result is a surprise for me for by the time the experiment is done these assumptions have been forgotten. These admissions lay me open to lectures on the subject of my vicious ways, but do not throw any doubt on my credibility when I testify to the surprises I experience.

I do not expect this reply to silence my critic. He will probably say that such surprises are due to some creative mental act on my part, and reflect no credit on the machine. This leads us back to the argument from consciousness, and far from the idea of surprise. It is a line of argument we must consider closed, but it is perhaps worth remarking that the appreciation of something as surprising requires as much of a 'creative mental act' whether the surprising event originates from a man, a book, a machine or anything else.

The view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. This is the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it. It is a very useful assumption under many circumstances, but one too easily forgets that it is false. A natural consequence of doing so is that one then assumes that there is no virtue in the mere working out of consequences from data and general principles.

(7) *Argument from Continuity in the Nervous System.* The nervous system is certainly not a discrete–state machine. A small error in the information about the size of a nervous impulse impinging on a neuron, may make a large difference to the size of the outgoing impulse. It may be argued that, this being so, one cannot expect to be able to mimic the behaviour of the nervous system with a discrete-state system.

It is true that a discrete-state machine must be different from a continuous machine. But if we adhere to the conditions of the imitation game, the interrogator will not be able to take any advantage of this difference. The situation can be made clearer if we consider some other simpler continuous machine. A differential analyser will do very well. (A differential analyser is a certain kind of machine not of the discrete-state type used for some kinds of calculation.) Some of these provide their answers in a typed form, and so are suitable for taking part in the game. It would not be possible for a digital computer to predict exactly what answers the differential analyser would give to a problem, but it would be quite capable of giving the right sort of answer. For instance, if asked to give the value of π (actually about 3·1416) it would be reasonable to choose at random between the values 3·12, 3·13, 3·14, 3·15, 3·16 with the probabilities of 0·05, 0·15, 0·55, 0·19, 0·06 (say). Under these circumstances it would be very difficult for the interrogator to distinguish the differential analyser from the digital computer.

(8) *The Argument from Informality of Behaviour.* It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances. One might for instance have a rule that one is to stop when one sees a red traffic light, and to go if one sees a green one, but what if by some fault both appear together? One may perhaps decide that it is safest to stop. But some further difficulty may well arise from this decision later. To attempt to provide rules of conduct to cover every eventuality, even those arising from traffic lights, appears to be impossible. With all this I agree.

From this it is argued that we cannot be machines. I shall try to reproduce the argument, but I fear I shall hardly do it justice. It seems to run something like this. 'If each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines.' The undistributed middle is glaring. I do not think the argument is ever put quite like this, but I believe this is the argument used nevertheless. There may however be a certain confusion between 'rules of conduct' and 'laws of behaviour' to cloud the issue. By 'rules of conduct' I mean precepts such as 'Stop if you see red lights' on which one can act, and of which one can be conscious. By 'laws of behaviour' I mean laws of nature as applied to a man's body such as 'if you pinch him he will squeak.' If we substitute 'laws of behaviour which regulate his life' for 'laws of conduct by which he regulates his life' in the argument quoted the undistributed middle is no longer insuperable. For we believe that it is not only true that being regulated by laws of behaviour implies being some sort of machine (though not necessarily a discrete-state machine), but that conversely being such a machine implies being regulated by such laws. However, we cannot so easily convince ourselves of the absence of complete laws of behaviour as of complete rules of conduct. The only way we know of for finding such laws is scientific observation, and we certainly know of no circumstances under which we could say, 'We have searched enough. There are no such laws.'

We can demonstrate more forcibly that any such statement would be unjustified. For suppose we could be sure of finding such laws if they existed. Then given a discrete-state machine it should certainly be possible to discover by observation sufficient about it to predict its future behaviour, and this within a reasonable time, say a thousand years. But this does not seem to be the case. I have set up on the Manchester computer a small programme using only 1,000 units of storage, whereby the machine supplied with one sixteen-figure number replies with another within two seconds. I would defy anyone to learn from these replies sufficient about the programme to be able to predict any replies to untried values.

(9) *The Argument from Extra-Sensory Perception.* I assume that the reader is familiar with the idea of extra-sensory perception, and the meaning of the four items of it, *viz.* telepathy, clairvoyance, precognition and psycho-kinesis. These disturbing phenomena seem to deny all our usual scientific ideas. How we should like to discredit them! Unfortunately the statistical evidence, at least for telepathy, is overwhelming. It is very difficult to rearrange one's ideas so as to fit these new facts in. Once one has accepted them it does not seem a very big step to believe in ghosts and bogies. The idea that our bodies move simply according to the known laws of physics, together with some others not yet discovered but somewhat similar, would be one of the first to go.

This argument is to my mind quite a strong one. One can say in reply that many scientific theories seem to remain workable in practice, in spite of clashing with E.S.P.; that in fact one can get along very nicely if one forgets about it. This is rather cold comfort, and one fears that

thinking is just the kind of phenomenon where E.S.P. may be especially, relevant.

A more specific argument based on E.S.P. might run as follows: 'Let us play the imitation game, using as witnesses a man who is good as a telepathic receiver, and a digital computer. The interrogator can ask such questions as 'What suit does the card in my right hand belong to?' The man by telepathy or clairvoyance gives the right answer 130 times out of 400 cards. The machine can only guess at random, and perhaps gets 104 right, so the interrogator makes the right identification.' There is an interesting possibility which opens here. Suppose the digital computer contains a random number generator. Then it will be natural to use this to decide what answer to give. But then the random number generator will be subject to the psycho-kinetic powers of the interrogator. Perhaps this psycho-kinesis might cause the machine to guess right more often than would be expected on a probability calculation, so that the interrogator might still be unable to make the right identification. On the other hand, he might be able to guess right without any questioning, by clairvoyance. With E.S.P. anything may happen.

If telepathy is admitted it will be necessary to tighten our test up. The situation could be regarded as analogous to that which would occur if the interrogator were talking to himself and one of the competitors was listening with his ear to the wall. To put the competitors into a 'telepathy-proof room' would satisfy all requirements.

7. Learning Machines

The reader will have anticipated that I have no very convincing arguments of a positive nature to support my views. If I had I should not have taken such pains to point out the fallacies in contrary views. Such evidence as I have I shall now give.

Let us return for a moment to Lady Lovelace's objection, which stated that the machine can only do what we tell it to do. One could say that a man can 'inject' an idea into the machine, and that it will respond to a certain extent and then drop into quiescence, like a piano string struck by a hammer. Another simile would be an atomic pile of less than critical size: an injected idea is to correspond to a neutron entering the pile from without. Each such neutron will cause a certain disturbance which eventually dies away. If, however, the size of the pile is sufficiently increased, the disturbance caused by such an incoming neutron will very likely go on and on increasing until the whole pile is destroyed. Is there a corresponding phenomenon for minds, and is there one for machines? There does seem to be one for the human mind. The majority of them seem to be 'sub-critical' *i.e.* to correspond in this analogy to piles of sub-critical size. An idea presented to such a mind will on average give rise to less than one idea in reply. A smallish proportion are super-critical. An idea presented to such a mind may give rise to a whole 'theory' consisting of secondary, tertiary and more remote ideas. Animals minds seem to be very definitely sub-critical. Adhering to this analogy we ask, 'Can a machine be made to be super-critical?'

The 'skin of an onion' analogy is also helpful. In considering the functions of the mind or the brain we find certain operations which we can explain in purely mechanical terms. This we say does not correspond to the real mind: it is a sort of skin which we must strip off if we are to find the real mind. But then in what remains we find a further skin to be stripped off, and so on.

Proceeding in this way do we ever come to the 'real' mind, or do we eventually come to the skin which has nothing in it? In the latter case the whole mind is mechanical. (It would not be a discrete-state machine however. We have discussed this.)

These last two paragraphs do not claim to be convincing arguments. They should rather be described as 'recitations tending to produce belief.'

The only really satisfactory support that can be given for the view expressed at the beginning of section 6, will be that provided by waiting for the end of the century and then doing the experiment described. But what can we say in the meantime? What steps should be taken now if the experiment is to be successful?

As I have explained, the problem is mainly one of programming. Advances in engineering will have to be made too, but it seems unlikely that these will not be adequate for the requirements. Estimates of the storage capacity of the brain vary from 10^{10} to 10^{15} binary digits. I incline to the lower values and believe that only a very small fraction is used for the higher types of thinking. Most of it is probably used for the retention of visual impressions. I should be surprised if more than 10^{9} was required for satisfactory playing of the imitation game, at any rate against a blind man. (Note—The capacity

of the *Encyclopaedia Britannica*, 11th edition, is 2×10^9.) A storage capacity of 10^7 would be a very practicable possibility even by present techniques. It is probably not necessary to increase the speed of operations of the machines at all. Parts of modern machines which can be regarded as analogues of nerve cells work about a thousand times faster than the latter. This should provide a 'margin of safety' which could cover losses of speed arising in many ways. Our problem then is to find out how to programme these machines to play the game. At my present rate of working I produce about a thousand digits of programme a day, so that about sixty workers, working steadily through the fifty years might accomplish the job, if nothing went into the waste-paper basket. Some more expeditious method seems desirable.

In the process of trying to imitate an adult human mind we are bound to think a good deal about the process which has brought it to the state that it is in. We may notice three components,

(*a*) The initial state of the mind, say at birth,
(*b*) The education to which it has been subjected,
(*c*) Other experience, not to be described as education, to which it has been subjected.

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

We have thus divided our problem into two parts. The child-programme and the education process. These two remain very closely connected. We cannot expect to find a good child-machine at the first attempt. One must experiment with teaching one such machine and see how well it teams. One can then try another and see if it is better or worse. There is an obvious connection between this process and evolution, by the identifications

Structure of the child machine = Hereditary material
Changes of the child-machine = Mutations
Natural selection = Judgment of the experimenter

One may hope, however, that this process will be more expeditious than evolution. The survival of the fittest is a slow method for measuring advantages. The experimenter, by the exercise of intelligence, should be able to speed it up. Equally important is the fact that he is not restricted to random mutations. If he can trace a cause for some weakness he can probably think of the kind of mutation which will improve it.

It will not be possible to apply exactly the same teaching process to the machine as to a normal child. It will not, for instance, be provided with legs, so that it could not be asked to go out and fill the coal scuttle. Possibly it might not have eyes. But however well these deficiencies might be overcome by clever engineering, one could not send the creature to school without the other children making excessive fun of it. It must be given some tuition. We need not be too concerned about the legs, eyes, etc. The example of Miss Helen Keller shows that education can take place provided that communication in both directions between teacher and pupil can take place by some means or other.

We normally associate punishments and rewards with the teaching process. Some simple child-machines can be constructed or programmed on this sort of principle. The machine has to be so constructed that events which shortly preceded the occurrence of a punishment-signal are unlikely to be repeated, whereas a reward-signal increased the probability of repetition of the events which led up to it. These definitions do not presuppose any feelings on the part of the machine. I have done some experiments with one such child-machine, and succeeded in teaching it a few things, but the teaching method was too unorthodox for the experiment to be considered really successful.

The use of punishments and rewards can at best be a part of the teaching process. Roughly speaking, if the teacher has no other means of communicating to the pupil, the amount of information which can reach him does not exceed the total number of rewards and punishments applied. By the time a child has learnt to repeat 'Casabianca' he would probably feel very sore indeed, if the text could only be discovered by a 'Twenty Questions' technique, every 'NO' taking the form of a blow. It is necessary

therefore to have some other 'unemotional' channels of communication. If these are available it is possible to teach a machine by punishments and rewards to obey orders given in some language, *e.g.* a symbolic language. These orders are to be transmitted through the 'unemotional' channels. The use of this language will diminish greatly the number of punishments and rewards required.

Opinions may vary as to the complexity which is suitable in the child machine. One might try to make it as simple as possible consistently with the general principles. Alternatively one might have a complete system of logical inference 'built in.'[3] In the latter case the store would be largely occupied with definitions and propositions. The propositions would have various kinds of status, *e.g.* well-established facts, conjectures, mathematically proved theorems, statements given by an authority, expressions having the logical form of proposition but not belief-value. Certain propositions may be described as 'imperatives.' The machine should be so constructed that as soon as an imperative is classed as 'well-established' the appropriate action automatically takes place. To illustrate this, suppose the teacher says to the machine, 'Do your homework now.' This may cause 'Teacher says 'Do your homework now' 'to be included amongst the well-established facts. Another such fact might be,' Everything that teacher says is true.' Combining these may eventually lead to the imperative, 'Do your homework now' being included amongst the well-established facts, and this, by the construction of the machine, will mean that the homework actually gets started, but the effect is very satisfactory. The processes of inference used by the machine need not be such as would satisfy the most exacting logicians. There might for instance be no hierarchy of types. But this need not mean that type fallacies will occur, any more than we are bound to fall over unfenced cliffs. Suitable imperatives (expressed *within* the systems, not forming part of the rules *of* the system) such as 'Do not use a class unless it is a subclass of one which has been mentioned by teacher' can have a similar effect to 'Do not go too near the edge.'

The imperatives that can be obeyed by a machine that has no limbs are bound to be of a rather intellectual character, as in the example (doing homework) given above. Important amongst such imperatives will be ones which regulate the order in which the rules of the logical system concerned are to be applied. For at each stage when one is using a logical system, there is a very large number of alternative steps, any of which one is permitted to apply, so far as obedience to the rules of the logical system is concerned. These choices make the difference between a brilliant and a footling reasoner, not the difference between a sound and a fallacious one. Propositions leading to imperatives of this kind might be 'When Socrates is mentioned, use the syllogism in Barbara' or 'If one method has been proved to be quicker than another, do not use the slower method.' Some of these may be 'given by authority' but others may be produced by the machine itself, *e.g.* by scientific induction.

The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. The reader may draw a parallel with the Constitution of the United States.

An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he may still be able to some extent to predict his pupil's behaviour. This should apply most strongly to the later education of a machine arising from a child-machine of well-tried design (or programme). This is in clear contrast with normal procedure when using a machine to do computations: one's object is then to have a clear mental picture of the state of the machine at each moment in the computation. This object can only be achieved with a struggle. The view that 'the machine can only do what we know how to order it to do'[4] appears strange in face of this. Most of the programmes which we can put into the machine will result in its doing something that we cannot make sense of at all, or which we regard as completely random behaviour. Intelligent behaviour presumably consists in a departure from the completely disciplined behaviour involved in computation, but a rather slight one, which does not give rise to random behaviour, or to pointless repetitive loops. Another important result of preparing our machine for its part in the imitation game by a process of teaching and learning is that 'human fallibility' is likely to be omitted in a rather natural way, *i.e.* without special 'coaching.'

Processes that are learnt do not produce a hundred per cent, certainty of result; if they did they could not be unlearnt.

It is probably wise to include a random element in a learning machine (see p. 438). A random element is rather useful when we are searching for a solution of some, problem. Suppose for instance we wanted to find a number between 50 and 200 which was equal to the square of the sum of its digits, we might start at 51 then try 52 and go on until we got a number that worked. Alternatively we might choose numbers at random until we got a good one. This method has the advantage that it is unnecessary to keep track of the values that have been tried, but the disadvantage that one may try the same one twice, but this is not very important if there are several solutions. The systematic method has the disadvantage that there may be an enormous block without any solutions in the region which has to be investigated first. Now the learning process may be regarded as a search for a form of behaviour which will satisfy the teacher (or some other criterion). Since there is probably a very large number of satisfactory solutions the random method seems to be better than the systematic. It should be noticed that it is used in the analogous process of evolution. But there the systematic method is not possible. How could one keep track of the different genetical combinations that had been tried, so as to avoid trying them again?

We may hope that machines will eventually compete with men in all purely intellectual fields. Bat which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

BIBLIOGRAPHY

Butler, Samuel, *The Book of the Machines* (London: Erewhon, 1865), chapters 23—25.

Church, Alonzo, "An Unsolvable Problem of Elementary Number Theory," *American Journal of Mathematics* 58 (1936): pp. 345–63.

Gödel, K., "Über formal unentscheidbare Sätzeder Principia Mathematica und verwandter Systeme, I," *Monatshefte für Math, und Phys* (1931): pp. 173–89.

Hartree, D. R. *Calculating Instruments and Machines* (New York, 1949).

Kleene, S. C., "General Recursive Functions of Natural Numbers," *American Journal of Mathematics*, 57 (1935): pp. 153–73 and 219–44.

Jefferson, G., "The Mind of Mechanical Man," (Lister Oration for 1949) *British Medical Journal*, vol. i (1949): pp. 1105–21.

Countess of Lovelace, "Translator's notes to an article on Babbage's Analytical Engine," in *Scientific Memoirs*, vol. 3, R. Taylor, ed., (1842), pp. 691–731.

Russell, Bertrand *History of Western Philosophy* (London:, 1940).

Turing, A. M., "On Computable Numbers, with an Application to the Entscheidungs Problem," *Proceedings of London Mathematical Society* 42, no. 2 (1937): pp. 230–65.

NOTES

1. Possibly this view is heretical. St. Thomas Aquinas (*Summa Theologica*, quoted by Bertrand Russell, p. 480) states that God cannot make a man to have no soul. But this may not be a real restriction on His powers, but only a result of the fact that men's souls are immortal, and therefore indestructible.
2. Authors' names in italics refer to the bibliography.
3. Or rather 'programmed in' for our child-machine will be programmed in a digital computer. But the logical system will not have to be learnt.
4. Compare Lady Lovelace's statement (p. 490), which does not contain the word 'only.'

Minds, Brains, and Programs

John R. Searle

What psychological and philosophical significance should we attach to recent efforts at computer simulations of human cognitive capacities? In answering this question, I find it useful to distinguish what I will call 'strong' AI from 'weak' or 'cautious' AI (Artificial Intelligence). According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations.

I have no objection to the claims of weak AI, at least as far as this article is concerned. My discussion here will be directed at the claims I have defined as those of strong AI, specifically the claim that the appropriately programmed computer literally has cognitive states and that the programs thereby explain human cognition. When I hereafter refer to AI, I have in mind the strong version, as expressed by these two claims.

I will consider the work of Roger Schank and his colleagues at Yale (Schank and Abelson 1977), because I am more familiar with it than I am with any other similar claims, and because it provides a very clear example of the sort of work I wish to examine. But nothing that follows depends upon the details of Schank's programs. The same arguments would apply to Winograd's SHRDLU (Winograd 1973), Weizenbaum's ELIZA (Weizenbaum 1965), and indeed any Turing machine simulation of human mental phenomena.

Very briefly, and leaving out the various details, one can describe Schank's program as follows: the aim of the program is to simulate the human ability to understand stories. It is characteristic of human beings' story—understanding capacity that they can answer questions about the story even though the information that they give was never explicitly stated in the story. Thus, for example, suppose you are given the following story: 'A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip.' Now, if you are asked 'Did the man eat the hamburger?' you will presumably answer, 'No, he did not.' Similarly, if you are given the following story: 'A man went into a restaurant and ordered a hamburger; when the hamburger came he was very pleased with it; and as he left the restaurant he gave the waitress a large tip before paying his bill,' and you are asked the question, 'Did the man eat the hamburger?,' you will presumably answer, 'Yes, he ate the hamburger.' Now Schank's machines can similarly answer questions about restaurants in this fashion. To do this, they have a 'representation' of the sort of information that human beings have about restaurants, which enables them to answer such questions as those above, given these sorts of stories. When the machine is given the story and then asked the question, the machine will print out answers of the sort that we would expect human beings to give if told similar stories. Partisans of strong AI claim that in this question and answer sequence the machine is not only simulating a human ability but also

1. that the machine can literally be said to *understand* the story and provide the answers to questions, and
2. that what the machine and its program do *explains* the human ability to understand the story and answer questions about it.

Both claims seem to me to be totally unsupported by Schank's[1] work, as I will attempt to show in what follows.

One way to test any theory of the mind is to ask oneself what it would be like if my mind

From *Behavioral and Brain Sciences* 3: 417–424, 1980. Reprinted by permission of the publisher.

actually worked on the principles that the theory says all minds work on. Let us apply this test to the Schank program with the following *Gedankenexperiment*. Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch 'a script,' they call the second batch a 'story,' and they call the third batch 'questions.' Furthermore, they call the symbols I give them back in response to the third batch 'answers to the questions,' and the set of rules in English that they gave me, they call 'the program.' Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from the point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view—from the point of view of someone reading my 'answers'—the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program.

Now the claims made by strong AI are that the programmed computer understands the stories and that the program in some sense explains human understanding. But we are now in a position to examine these claims in light of our thought experiment.

1. As regards the first claim, it seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. For the same reasons, Schank's computer understands nothing of any stories, whether in Chinese, English, or whatever, since in the Chinese case the computer is me, and in cases where the computer is not me, the computer has nothing more than I have in the case where I understand nothing.
2. As regards the second claim, that the program explains human understanding, we can see that the computer and its program do not provide sufficient conditions of understanding since the computer and the program are functioning, and there is no understanding. But does it even provide a necessary condition or a significant contribution to understanding? One of the claims made by the supporters of strong AI is that when I understand a story in English, what I am doing is exactly the same—or perhaps more of the same—as what I was doing in manipulating the Chinese symbols. It is simply more formal symbol manipulation that distinguishes the case in English, where I do understand, from the case in Chinese, where I don't. I have not demonstrated that this claim is false, but it would certainly appear an incredible claim in the example. Such plausibility as the claim has derives from the

supposition that we can construct a program that will have the same inputs and outputs as native speakers, and in addition we assume that speakers have some level of description where they are also instantiations of a program. On the basis of these two assumptions we assume that even if Schank's program isn't the whole story about understanding, it may be part of the story. Well, I suppose that is an empirical possibility, but not the slightest reason has so far been given to believe that it is true, since what is suggested—though certainly not demonstrated—by the example is that the computer program is simply irrelevant to my understanding of the story. In the Chinese case I have everything that artificial intelligence can put into me by way of a program, and I understand nothing; in the English case I understand everything, and there is so far no reason at all to suppose that my understanding has anything to do with computer programs, that is, with computational operations on purely formally specified elements. As long as the program is defined in terms of computational operations on purely formally defined elements, what the example suggests is that these by themselves have no interesting connection with understanding. They are certainly not sufficient conditions, and not the slightest reason has been given to suppose that they are necessary conditions or even that they make a significant contribution to understanding. Notice that the force of the argument is not simply that different machines can have the same input and output while operating on different formal principles—that is not the point at all. Rather, whatever purely formal principles you put into the computer, they will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything. No reason whatever has been offered to suppose that such principles are necessary or even contributory, since no reason has been given to suppose that when I understand English I am operating with any formal program at all.

Well, then, what is it that I have in the case of the English sentences that I do not have in the case of the Chinese sentences? The obvious answer is that I know what the former mean, while I haven't the faintest idea what the latter mean. But in what does this consist and why couldn't we give it to a machine, whatever it is? I will return to this question later, but first I want to continue with the example.

I have had the occasions to present this example to several workers in artificial intelligence, and, interestingly, they do not seem to agree on what the proper reply to it is. I get a surprising variety of replies, and in what follows I will consider the most common of these (specified along with their geographic origins).

But first I want to block some common misunderstandings about 'understanding'; in many of these discussions one finds a lot of fancy footwork about the word 'understanding.' My critics point out that there are many different degrees of understanding; that 'understanding' is not a simple two-place predicate; that there are even different kinds and levels of understanding, and often the law of excluded middle doesn't even apply in a straightforward way to statements of the form 'x understands y'; that in many cases it is a matter for decision and not a simple matter of fact whether x understands y; and so on. To all of these points I want to say: of course, of course. But they have nothing to do with the points at issue. There are clear cases in which 'understanding' literally applies and clear cases in which it does not apply; and these two sorts of cases are all I need for this argument.[2] I understand stories in English; to a lesser degree I can understand stories in French; to a still lesser degree, stories in German; and in Chinese, not at all. My car and my adding machine, on the other hand, understand nothing; they are not in that line of business. We often attribute 'understanding' and other cognitive predicates by metaphor and analogy to cars, adding machines, and other artifacts, but nothing is proved by such attributions. We say, 'The door *blows* when to open because of its photoelectric cell,' 'The adding machine *knows how* (*understands how*, is *able*) to do addition and subtraction but not division,' and 'The thermostat *perceives* chances in the temperature.' The reason we make these attributions is quite interesting, and it has to do with the fact that in artifacts we extend our own intentionally;[3] our tools are extensions of our purposes, and so we find it natural to make metaphorical attributions of intentionality to them; but I take it no philosophical ice is cut by such examples. The sense in which an automatic door 'understands instructions' from its photoelectric cell is not at all the sense in which I understand English. If the sense in which Schank's programmed computers understand stories is supposed to be the

metaphorical sense in which the door understands, and not the sense in which I understand English, the issue would not be worth discussing. But Newell and Simon 1963 write that the kind of cognition they claim for computers is exactly the same as for human beings. I like the straightforwardness of this claim, and it is the sort of claim I will be considering. I will argue that in the literal sense the programmed computer understands what the car and the adding machine understand, namely, exactly nothing. The computer understanding is not just (like my understanding of German) partial or incomplete; it is zero.

Now to the replies:

I. The systems reply (Berkeley).

'While it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand the story. The person has a large ledger in front of him in which are written the rules, he has a lot of scratch paper and pencils for doing calculations, he has 'data banks' of sets of Chinese symbols. Now, understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part.'

My response to the systems theory is quite simple; let the individual internalize all of these elements of the system. He memorizes the rules in the ledger and the data banks of Chinese symbols, and he does all the calculations in his head. The individual then incorporates the entire system. There isn't anything at all to the system that he does not encompass. We can even get rid of the room and suppose he works outdoors. All the same, he understands nothing of the Chinese, and a fortiori neither does the system, because there isn't anything in the system that isn't in him. If he doesn't understand, then there is no way the system could understand because the system is just a part of him.

Actually I feel somewhat embarrassed to give even this answer to the systems theory because the theory seems to me so implausible to start with. The idea is that while a person doesn't understand Chinese, somehow the *conjunction* of that person and bits of paper might understand Chinese. It is not easy for me to imagine how someone who was not in the grip of an ideology would find the idea at all plausible. Still, I think many people who are committed to the ideology of strong AI will in the end be inclined to say something very much like this; so let us pursue it a bit further. According to one version of this view, while the man in the internalized systems example doesn't understand Chinese in the sense that a native Chinese speaker does (because, for example, he doesn't know that the story refers to restaurants and hamburgers, etc.), still 'the man as a formal symbol manipulation system' *really does understand Chinese*. The subsystem of the man that is the formal symbol manipulation system for Chinese should not be confused with the subsystem for English.

So there are really two subsystems in the man; one understands English, the other Chinese, and 'it's just that the two systems have little to do with each other.' But, I want to reply, not only do they have little to do with each other, they are not even remotely alike. The subsystem that understands English (assuming we allow ourselves to talk in this jargon of 'subsystems' for a moment) knows that the stories are about restaurants and eating hamburgers, he knows that he is being asked questions about restaurants and that he is answering questions as best he can by making various inferences from the content of the story, and so on. But the Chinese system knows none of this. Whereas the English subsystem knows that 'hamburgers' refers to hamburgers, the Chinese subsystem knows only that 'squiggle squiggle' is followed by 'squoggle squoggle.' All he knows is that various formal symbols are being introduced at one end and manipulated according to rules written in English, and other symbols are going out at the other end. The whole point of the original example was to argue that such symbol manipulation by itself couldn't be sufficient for understanding Chinese in any literal sense because the man could write 'squoggle squoggle' after 'squiggle squiggle' without understanding anything in Chinese. And it doesn't meet that argument to postulate subsystems within the man, because the subsystems are no better off than the man was in the first place; they still don't have anything even remotely like what the English-speaking man (or subsystem) has. Indeed, in the case as described, the Chinese subsystem is simply a part of the English subsystem, a part that engages in meaningless symbol manipulation according to rules in English.

Let us ask ourselves what is supposed to motivate the systems reply in the first place; that is, what *independent* grounds are there supposed to be for saying that the agent must have a subsystem within him that literally understands

stories in Chinese? As far as I can tell the only grounds are that in the example I have the same input and output as native Chinese speakers and a program that goes from one to the other. But the whole point of the examples has been to try to show that that couldn't be sufficient for understanding, in the sense in which I understand stories in English, because a person, and hence the set of systems that go to make up a person, could have the right combination of input, output, and program and still not understand anything in the relevant literal sense in which I understand English. The only motivation for saying there *must* be a subsystem in me that understands Chinese is that I have a program and I can pass the Turing test; I can fool native Chinese speakers. But precisely one of the points at issue is the adequacy of the Turing test. The example shows that there could be two 'systems,' both of which pass the Turing test, but only one of which understands; and it is no argument against this point to say that since they both pass the Turing test they must both understand, since this claim fails to meet the argument that the system in me that understands English has a great deal more than the system that merely processes Chinese. In short, the systems reply simply begs the question by insisting without argument that the system must understand Chinese.

Furthermore, the systems reply would appear to lead to consequences that are independently absurd. If we are to conclude that there must be cognition in me on the grounds that I have a certain sort of input and output and a program in between, then it looks like all sorts of noncognitive subsystems are going to turn out to be cognitive. For example, there is a level of description at which my stomach does information processing, and it instantiates any number of computer programs, but I take it we do not want to say that it has any understanding [cf. Pylyshyn, "Computation and Cognitition" *BBS* 3, no. 1 (1980)]. But if we accept the systems reply, then it is hard to see how we avoid saying that stomach, heart, liver, and so on, are all understanding subsystems, since there is no principled way to distinguish the motivation for saying the Chinese subsystem understands from saying that the stomach understands. It is, by the way, not an answer to this point to say that the Chinese system has information as input and output and the stomach has food and food products as input and output, since from the point of view of the agent, from my point of view, there is no information in either the food or the Chinese—the Chinese is just so many meaningless squiggles. The information in the Chinese case is solely in the eyes of the programmers and the interpreters, and there is nothing to prevent them from treating the input and output of my digestive organs as information if they so desire.

This last point bears on some independent problems in strong AI, and it is worth digressing for a moment to explain it. If strong AI is to be a branch of psychology, then it must be able to distinguish those systems that are genuinely mental from those that are not. It must be able to distinguish the principles on which the mind works from those on which nonmental systems work; otherwise it will offer us no explanations of what is specifically mental about the mental. And the mental-nonmental distinction cannot be just in the eye of the beholder but it must be intrinsic to the systems; otherwise it would be up to any beholder to treat people as nonmental and, for example, hurricanes as mental if he likes. But quite often in the AI literature the distinction is blurred in ways that would in the long run prove disastrous to the claim that AI is a cognitive inquiry. McCarthy, for example, writes, 'Machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines capable of problem solving performance' (McCarthy 1979). Anyone who thinks strong AI has a chance as a theory of the mind ought to ponder the implications of that remark. We are asked to accept it as a discovery of strong AI that the hunk of metal on the wall that we use to regulate the temperature has beliefs in exactly the same sense that we, our spouses, and our children have beliefs, and furthermore that 'most' of the other machines in the room—telephone, tape recorder, adding machine, electric light switch,—also have beliefs in this literal sense. It is not the aim of this article to argue against McCarthy's point, so I will simply assert the following without argument. The study of the mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don't. If you get a theory that denies this point you have produced a counterexample to the theory and the theory is false. One gets the impression that people in AI who write this sort of thing think they can get away with it because they don't really take it seriously, and they don't think anyone else will either. I propose for a moment at least, to take it seriously. Think hard for one minute about what would be necessary

to establish that that hunk of metal on the wall over there had real beliefs, beliefs with direction of fit, propositional content, and conditions of satisfaction; beliefs that had the possibility of being strong beliefs or weak beliefs; nervous, anxious, or secure beliefs; dogmatic, rational, or superstitious beliefs; blind faiths or hesitant cogitations; any kind of beliefs. The thermostat is not a candidate. Neither is stomach, liver, adding machine, or telephone. However, since we are taking the idea seriously, notice that its truth would be fatal to strong AI's claim to be a science of the mind. For now the mind is everywhere. What we wanted to know is what distinguishes the mind from thermostats and livers. And if McCarthy were right, strong AI wouldn't have a hope of telling us that.

II. The robot reply (Yale).

'Suppose we wrote a different kind of program from Schank's program. Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking—anything you like. The robot would, for example, have a television camera attached to it that enabled it to 'see,' it would have arms and legs that enabled it to 'act,' and all of this would be controlled by its computer 'brain. 'Such a robot would, unlike Schank's computer, have genuine understanding and other mental states.'

The first thing to notice about the robot reply is that it tacitly concedes that cognition is not solely a matter of formal symbol manipulation, since this reply adds a set of causal relation with the outside world [cf. Fodor, "Methodological Solipsism," *BBS* 3, no. 1 (1980)]. But the answer to the robot reply is that the addition of such 'perceptual' and 'motor' capacities adds nothing by way of understanding, in particular, or intentionality, in general, to Schank's original program. To see this, notice that the same thought experiment applies to the robot case. Suppose that instead of the computer inside the robot, you put me inside the room and, as in the original Chinese case, you give me more Chinese symbols with more instructions in English for matching Chinese symbols to Chinese symbols and feeding back Chinese symbols to the outside. Suppose, unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot's legs or arms. It is important to emphasize that all I am doing is manipulating formal symbols: I know none of these other facts. I am receiving 'information' from the robot's 'perceptual' apparatus, and I am giving out 'instructions' to its motor apparatus without knowing either of these facts. I am the robot's homunculus, but unlike the traditional homunculus, I don't know what's going on. I don't understand anything except the rules for symbol manipulation. Now in this case I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. And furthermore, by instantiating the program I have no intentional states of the relevant type. All I do is follow formal instructions about manipulating formal symbols.

III. The brain simulator reply (Berkeley and M.I.T.).

'Suppose we design a program that doesn't represent information that we have about the world, such as the information in Schank's scripts, but simulates the actual sequence of neuron firings at the synapses of the brain of a native Chinese speaker when he understands stories in Chinese and gives answers to them. The machine takes in Chinese stories and questions about them as input, it simulates the formal structure of actual Chinese brains in processing these stories, and it gives out Chinese answers as outputs. We can even imagine that the machine operates, not with a single serial program, but with a whole set of programs operating in parallel, in the manner that actual human brains presumably operate when they process natural language. Now surely in such a case we would have to say that the machine understood the stories; and if we refuse to say that, wouldn't we also have to deny that native Chinese speakers understood the stories? At the level of the synapses, what would or could be different about the program of the computer and the program of the Chinese brain?'

Before countering this reply I want to digress to note that it is an odd reply for any partisan of artificial intelligence (or functionalism, etc.) to make: I thought the whole idea of strong AI is that we don't need to know how the brain works to know how the mind works. The basic hypothesis, or so I had supposed, was that there is a

level of mental operations consisting of computational processes over formal elements that constitute the essence of the mental and can be realized in all sorts of different brain processes, in the same way that any computer program can be realized in different computer hardwares: on the assumptions of strong AI, the mind is to the brain as the program is to the hardware, and thus we can understand the mind without doing neurophysiology. If we had to know how the brain worked to do AI, we wouldn't bother with AI. However, even getting this close to the operation of the brain is still not sufficient to produce understanding. To see this, imagine that instead of a monolingual man in a room shuffling symbols we have the man operate an elaborate set of water pipes with valves connecting them. When the man receives the Chinese symbols, he looks up in the program, written in English, which valves he has to turn on and off. Each water connection corresponds to a synapse in the Chinese brain, and the whole system is rigged up so that after doing all the right firings, that is after turning on all the right faucets, the Chinese answers pop out at the output end of the series of pipes.

Now where is the understanding in this system? It takes Chinese as input, it simulates the formal structure of the synapses of the Chinese brain, and it gives Chinese as output. But the man certainly doesn't understand Chinese, and neither do the water pipes, and if we are tempted to adopt what I think is the absurd view that somehow the *conjunction* of man *and* water pipes understands, remember that in principle the man can internalize the formal structure of the water pipes and do all the 'neuron firings' in his imagination. The problem with the brain simulator is that it is simulating the wrong things about the brain. As long as it simulates only the formal structure of the sequence of neuron firings at the synapses, it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states. And that the formal properties are not sufficient for the causal properties is shown by the water pipe example: we can have all the formal properties carved off from the relevant neurobiological causal properties.

IV. The combination reply (Berkeley and Stanford).

'While each of the previous three replies might not be completely convincing by itself as a refutation of the Chinese room counterexample, if you take all three together they are collectively much more convincing and even decisive. Imagine a robot with a brain-shaped computer lodged in its cranial cavity, imagine the computer programmed with all the synapses of a human brain, imagine the whole behavior of the robot is indistinguishable from human behavior, and now think of the whole thing as a unified system and not just as a computer with inputs and outputs. Surely in such a case we would have to ascribe intentionality to the system.'

I entirely agree that in such a case we would find it rational and indeed irresistible to accept the hypothesis that the robot had intentionality, as long as we knew nothing more about it. Indeed, besides appearance and behavior, the other elements of the combination are really irrelevant. If we could build a robot whose behavior was indistinguishable over a large range from human behavior, we would attribute intentionality to it, pending some reason not to. We wouldn't need to know in advance that its computer brain was a formal analogue of the human brain.

But I really don't see that this is any help to the claims of strong AI; and here's why: According to strong AI, instantiating a formal program with the right input and output is a sufficient condition of, indeed is constitutive of, intentionality. As Newell 1979 puts it, the essence of the mental is the operation of a physical symbol system. But the attributions of intentionality that we make to the robot in this example have nothing to do with formal programs. They are simply based on the assumption that if the robot looks and behaves sufficiently like us, then we would suppose, until proven otherwise, that it must have mental states like ours that cause and are expressed by its behavior and it must have an inner mechanism capable of producing such mental states. If we knew independently how to account for its behavior without such assumptions we would not attribute intentionality to it, especially if we knew it had a formal program. And this is precisely the point of my earlier reply to objection II.

Suppose we knew that the robot's behavior was entirely accounted for by the fact that a man inside it was receiving uninterpreted formal symbols from the robot's sensory receptors and sending out uninterpreted formal symbols to its motor mechanisms, and the man was doing this symbol manipulation in accordance with a bunch of rules. Furthermore, suppose the man knows none of these facts about the robot, all he

knows is which operations to perform on which meaningless symbols. In such a case we would regard the robot as an ingenious mechanical dummy. The hypothesis that the dummy has a mind would now be unwarranted and unnecessary, for there is now no longer any reason to ascribe intentionality to the robot or to the system of which it is a part (except of course for the man's intentionality in manipulating the symbols). The formal symbol manipulations go on, the input and output are correctly matched, but the only real locus of intentionality is the man, and he doesn't know any of the relevant intentional states; he doesn't, for example, *see* what comes into the robot's eyes, he doesn't *intend* to move the robot's arm, and he doesn't *understand* any of the remarks made to or by the robot. Nor, for the reasons stated earlier, does the system of which man and robot are a part.

To see this point, contrast this case with cases in which we find it completely natural to ascribe intentionality to members of certain other primate species such as apes and monkeys and to domestic animals such as dogs. The reasons we find it natural are, roughly, two: we can't make sense of the animal's behavior without the ascription of intentionality, and we can see that the beasts are made of similar stuff to ourselves—that is an eye, that a nose, this is its skin, and so on. Given the coherence of the animal's behavior and the assumption of the same causal stuff underlying it, we assume both that the animal must have mental states underlying its behavior, and that the mental states must be produced by mechanisms made out of the stuff that is like our stuff. We would certainly make similar assumptions about the robot unless we had some reason not to, but as soon as we knew that the behavior was the result of a formal program, and that the actual causal properties of the physical substance were irrelevant we would abandon the assumption of intentionality. [See "Cognition and Consciousness in Nonhuman Species," *BBS* 1, no. 4 (1978).]

There are two other responses to my example that come up frequently (and so are worth discussing) but really miss the point.

V. The other minds reply (Yale).

'How do you know that other people understand Chinese or anything else? Only by their behavior. Now the computer can pass the behavioral tests as well as they can (in principle), so if you are going to attribute cognition to other people you must in principle also attribute it to computers.'

This objection really is only worth a short reply. The problem in this discussion is not about how I know that other people have cognitive states, but rather what it is that I am attributing to them when I attribute cognitive states to them. The thrust of the argument is that it couldn't be just computational processes and their output because the computational processes and their output can exist without the cognitive state. It is no answer to this argument to feign anesthesia. In 'cognitive sciences' one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects.

VI. The many mansions reply (Berkeley).

'Your whole argument presupposes that AI is only about analogue and digital computers. But that just happens to be the present state of technology. Whatever these causal processes are that you say are essential for intentionality (assuming you are right), eventually we will be able to build devices that have these causal processes, and that will be artificial intelligence. So your arguments are in no way directed at the ability of artificial intelligence to produce and explain cognition.'

I really have no objection to this reply save to say that it in effect trivializes the project of strong AI by redefining it as whatever artificially produces and explains cognition. The interest of the original claim made on behalf of artificial intelligence is that it was a precise, well defined thesis: mental processes are computational processes over formally defined elements. I have been concerned to challenge that thesis. If the claim is redefined so that it is no longer that thesis, my objections no longer apply because there is no longer a testable hypothesis for them to apply to.

Let us now return to the question I promised I would try to answer: granted that in my original example I understand the English and I do not understand the Chinese, and granted therefore that the machine doesn't understand either English or Chinese, still there must be something about me that makes it the case that I understand English and a corresponding something lacking in me that makes it the case that I

fail to understand Chinese. Now why couldn't we give those somethings, whatever they are, to a machine?

I see no reason in principle why we couldn't give a machine the capacity to understand English or Chinese, since in an important sense our bodies with our brains are precisely such machines. But I do see very strong arguments for saying that we could not give such a thing to a machine where the operation of the machine is defined solely in terms of computational processes over formally defined elements; that is, where the operation of the machine is defined as an instantiation of a computer program. It is not because I am the instantiation of a computer program that I am able to understand English and have other forms of intentionality (I am, I suppose, the instantiation of any number of computer programs), but as far as we know it is because I am a certain sort of organism with a certain biological (i.e. chemical and physical) structure, and this structure, under certain conditions, is causally capable of producing perception, action, understanding, learning, and other intentional phenomena. And part of the point of the present argument is that only something that had those causal powers could have that intentionality. Perhaps other physical and chemical processes could produce exactly these effects; perhaps, for example, Martians also have intentionality but their brains are made of different stuff. That is an empirical question, rather like the question whether photosynthesis can be done by something with a chemistry different from that of chlorophyll.

But the main point of the present argument is that no purely formal model will ever be sufficient by itself for intentionality because the formal properties are not by themselves constitutive of intentionality, and they have by themselves no causal powers except the power, when instantiated, to produce the next stage of the formalism when the machine is running. And any other causal properties that particular realizations of the formal model have, are irrelevant to the formal model because we can always put the same formal model in a different realization where those causal properties are obviously absent. Even if, by some miracle, Chinese speakers exactly realize Schank's program, we can put the same program in English speakers, water pipes, or computers, none of which understand Chinese, the program notwithstanding.

What matters about brain operations is not the formal shadow cast by the sequence of synapses but rather the actual properties of the sequences. All the arguments for the strong version of artificial intelligence that I have seen insist on drawing an outline around the shadows cast by cognition and then claiming that the shadows are the real thing.

By way of concluding I want to try to state some of the general philosophical points implicit in the argument. For clarity I will try to do it in a question and answer fashion, and I begin with that old chestnut of a question:

'Could a machine think?'

The answer is, obviously, yes. We are precisely such machines.

'Yes, but could an artifact, a man-made machine, think?'

Assuming it is possible to produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it, sufficiently like ours, again the answer to the question seems to be obviously, yes. If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those that human beings use. It is, as I said, an empirical question.

'OK, but could a digital computer think?'

If by 'digital computer' we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of course, yes, since we are the instantiations of any number of computer programs, and we can think.

'But could something think, understand, and so on *solely* in virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?'

This I think is the right question to ask, though it is usually confused with one or more of the earlier questions, and the answer to it is no.

'Why not?'

Because the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even *symbol* manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output.

The aim of the Chinese room example was to try to show this by showing that as soon as

we put something into the system that really does have intentionality (a man), and we program him with the formal program, you can see that the formal program carries no additional intentionality. It adds nothing, for example, to a man's ability to understand Chinese.

Precisely that feature of AI that seemed so appealing—the distinction between the program and the realization—proves fatal to the claim that simulation could be duplication. The distinction between the program and its realization in the hardware seems to be parallel to the distinction between the level of mental operations and the level of brain operations. And if we could describe the level of mental operations as a formal program, then it seems we could describe what was essential about the mind without doing either introspective psychology or neurophysiology of the brain. But the equation, 'mind is to brain as program is to hardware' breaks down at several points, among them the following three:

First, the distinction between program and realization has the consequence that the same program could have all sorts of crazy realizations that had no form of intentionality. Weizenbaum 1976, chapter 2, for example, shows in detail how to construct a computer using a roll of toilet paper and a pile of small stones. Similarly, the Chinese story understanding program can be programmed into a sequence of water pipes, a set of wind machines, or a monolingual English speaker, none of which thereby acquires an understanding of Chinese. Stones, toilet paper, wind, and water pipes are the wrong kind of stuff to have intentionality in the first place—only something that has the same causal powers as brains can have intentionality—and though the English speaker has the right kind of stuff for intentionality you can easily see that he doesn't get any extra intentionality by memorizing the program, since memorizing it won't teach him Chinese.

Second, the program is purely formal, but the intentional states are not in that way formal. They are defined in terms of their content, not their form. The belief that it is raining, for example, is not defined as a certain formal shape, but as a certain mental content with conditions of satisfaction, a direction of fit (see Searle 1979), and the like. Indeed the belief as such hasn't even got a formal shape in this syntactic sense, since one and the same belief can be given an indefinite number of different syntactic expressions in different linguistic systems.

Third, as I mentioned before, mental states and events are literally a product of the operation of the brain, but the program is not in that way a product of the computer.

'Well if programs are in no way constitutive of mental processes, why have so many people believed the converse? That at least needs some explanation.'

I don't really know the answer to that one. The idea that computer simulations could be the real thing ought to have seemed suspicious in the first place because the computer isn't confined to simulating mental operations, by any means. No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down or that a computer simulation of a rainstorm will leave us all drenched. Why on earth would anyone suppose that a computer simulation of understanding actually understood anything? It is sometimes said that it would be frightfully hard to get computers to feel pain or fall in love, but love and pain are neither harder nor easier than cognition or anything else. For simulation, all you need is the right input and output and a program in the middle that transforms the former into the latter. That is all the computer has for anything it does. To confuse simulation with duplication is the same mistake, whether it is pain, love, cognition, fires, or rainstorms.

Still, there are several reasons why AI must have seemed—and to many people perhaps still does seem—in some way to reproduce and thereby explain mental phenomena, and I believe we will not succeed in removing these illusions until we have fully exposed the reasons that give rise to them.

First, and perhaps most important, is a confusion about the notion of 'information processing': many people in cognitive science believe that the human brain, with its mind, does something called 'information processing,' and analogously the computer with its program does information processing; but fires and rainstorms, on the other hand, don't do information processing at all. Thus, though the computer can simulate the formal features of any process whatever, it stands in a special relation to the mind and brain because when the computer is properly programmed, ideally with the same program as the brain, the information processing is identical in the two cases, and this information processing is really the essence of the mental. But the trouble with this argument is that it rests on an ambiguity in the notion of 'information.' In the sense in which people 'process information' when they reflect, say, on problems in arithmetic or when they read and answer questions about stories, the programmed computer does not do 'information

processing.' Rather, what it does is manipulate formal symbols. The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. The computer, to repeat, has a syntax but no semantics. Thus, if you type into the computer '2 plus 2 equals?' it will type out '4.' But it has no idea that '4' means 4 or that it means anything at all. And the point is not that it lacks some second-order information about the interpretation of its first-order symbols, but rather that its first-order symbols don't have any interpretations as far as the computer is concerned. All the computer has is more symbols. The introduction of the notion of 'information processing' therefore produces a dilemma: either we construe the notion of 'information processing' in such a way that it implies intentionality as part of the process or we don't. If the former, then the programmed computer does not do information processing, it only manipulates formal symbols. If the latter, then, though the computer does information processing, it is only doing so in the sense in which adding machines, typewriters, stomachs, thermostats, rainstorms, and hurricanes do information processing; namely, they have a level of description at which we can describe them as taking information in at one end, transforming it, and producing information as output. But in this case it is up to outside observers to interpret the input and output as information in the ordinary sense. And no similarity is established between the computer and the brain in terms of any similarity of information processing.

Second, in much of AI there is a residual behaviorism or operationalism. Since appropriately programmed computers can have input-output patterns similar to those of human beings, we are tempted to postulate mental states in the computer similar to human mental states. But once we see that it is both conceptually and empirically possible for a system to have human capacities in some realm without having any intentionality at all, we should be able to overcome this impulse. My desk adding machine has calculating capacities, but no intentionality, and in this paper I have tried to show that a system could have input and output capabilities that duplicated those of a native Chinese speaker and still not understand Chinese, regardless of how it was programmed. The Turing test is typical of the tradition in being unashamedly behavioristic and operationalistic, and I believe that if AI workers totally repudiated behaviorism and operationalism much of the confusion between simulation and duplication would be eliminated.

Third, this residual operationalism is joined to a residual form of dualism; indeed strong AI only makes sense given the dualistic assumption that, where the mind is concerned, the brain doesn't matter. In strong AI (and in functionalism, as well) what matters are programs, and programs are independent of their realization in machines; indeed, as far as AI is concerned, the same program could be realized by an electronic machine, a Cartesian mental substance, or a Hegelian world spirit. The single most surprising discovery that I have made in discussing these issues is that many AI workers are quite shocked by my idea that actual human mental phenomena might be dependent on actual physical-chemical properties of actual human brains. But if you think about it a minute you can see that I should not have been surprised; for unless you accept some form of dualism, the strong AI project hasn't got a chance. The project is to reproduce and explain the mental by designing programs, but unless the mind is not only conceptually but empirically independent of the brain you couldn't carry out the project, for the program is completely independent of any realization. Unless you believe that the mind is separable from the brain both conceptually and empirically—dualism in a strong form—you cannot hope to reproduce the mental by writing and running programs since programs must be independent of brains or any other particular forms of instantiation. If mental operations consist in computational operations on formal symbols, then it follows that they have no interesting connection with the brain; the only connection would be that the brain just happens to be one of the indefinitely many types of machines capable of instantiating the program. This form of dualism is not the traditional Cartesian variety that claims there are two sorts of *substances*, but it is Cartesian in the sense that it insists that what is specifically mental about the mind has no intrinsic connection with the actual properties of the brain. This underlying dualism is masked from us by the fact that AI literature contains frequent fulminations against 'dualism'; what the authors seem to be unaware of is that their position presupposes a strong version of dualism.

'Could a machine think?' My own view is that *only* a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains. And that is the main reason strong AI has had little to tell us about thinking, since it has nothing to tell us about machines. By its own definition, it is about programs, and

programs are not machines. Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle because of a deep and abiding dualism: the mind they suppose is a matter of formal processes and is independent of quite specific material causes in the way that milk and sugar are not.

In defense of this dualism the hope is often expressed that the brain is a digital computer (early computers, by the way, were often called 'electronic brains'). But that is no help. Of course the brain is a digital computer. Since everything is a digital computer, brains are too. The point is that the brain's causal capacity to produce intentionality cannot consist in its instantiating a computer program, since for any program you like it is possible for something to instantiate that program and still not have any mental states. Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality.

ACKNOWLEDGMENTS

I am indebted to a rather large number of people for discussion of these matters and for their patient attempts to overcome my ignorance of artificial intelligence. I would especially like to thank Ned Block, Hubert Dreyfus, John Haugeland, Roger Schank, Robert Wilensky, and Terry Winograd.

NOTES

1. I am not, of course, saying that Schank himself is committed to these claims.
2. Also, 'understanding' implies both the possession of mental (intentional) states and the truth (validity, success) of these states. For the purposes of this discussion we are concerned only with the possession of the states.
3. Intentionality is by definition that feature of certain mental states by which they are directed at or about objects and states of affairs in the world. Thus, beliefs, desires, and intentions are intentional states; undirected forms of anxiety and depression are not. For further discussion see Searle 1979c.

The Singularity: A Philosophical Analysis

David J. Chalmers

1. Introduction[1]

What happens when machines become more intelligent than humans? One view is that this event will be followed by an explosion to ever-greater levels of intelligence, as each generation of machines creates more intelligent machines in turn. This intelligence explosion is now often known as the 'singularity.'

The basic argument here was set out by the statistician I. J. Good in his 1965 article "Speculations Concerning the First Ultraintelligent Machine":

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could

From *Journal of Consciousness Studies* 17:7-65, 2010. Reprinted by permission of the author.

> design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

The key idea is that a machine that is more intelligent than humans will be better than humans at designing machines. So it will be capable of designing a machine more intelligent than the most intelligent machine that humans can design. So if it is itself designed by humans, it will be capable of designing a machine more intelligent than itself. By similar reasoning, this next machine will also be capable of designing a machine more intelligent than itself. If every machine in turn does what it is capable of, we should expect a sequence of ever more intelligent machines.[2]

This intelligence explosion is sometimes combined with another idea, which we might call the 'speed explosion.' The argument for a speed explosion starts from the familiar observation that computer processing speed doubles at regular intervals. Suppose that speed doubles every two years and will do so indefinitely. Now suppose that we have human-level artificial intelligence designing new processors. Then faster processing will lead to faster designers and an ever-faster design cycle, leading to a limit point soon afterwards.

The argument for a speed explosion was set out by the artificial intelligence researcher Ray Solomonoff in his 1985 article "The Time Scale of Artificial Intelligence."[3] Eliezer Yudkowsky gives a succinct version of the argument in his 1996 article "Staring at the Singularity":

> 'Computing speed doubles every two subjective years of work. Two years after Artificial Intelligences reach human equivalence, their speed doubles. One year later, their speed doubles again. Six months—three months—1.5 months . . . Singularity.'

The intelligence explosion and the speed explosion are logically independent of each other. In principle there could be an intelligence explosion without a speed explosion and a speed explosion without an intelligence explosion. But the two ideas work particularly well together. Suppose that within two subjective years, a greater-than-human machine can produce another machine that is not only twice as fast but 10% more intelligent, and suppose that this principle is indefinitely extensible. Then within four objective years there will have been an infinite number of generations, with both speed and intelligence increasing beyond any finite level within a finite time. This process would truly deserve the name 'singularity.'

Of course the laws of physics impose limitations here. If the currently accepted laws of relativity and quantum mechanics are correct—or even if energy is finite in a classical universe—then we cannot expect the principles above to be indefinitely extensible. But even with these physical limitations in place, the arguments give some reason to think that both speed and intelligence might be pushed to the limits of what is physically possible. And on the face of it, it is unlikely that human processing is even close to the limits of what is physically possible. So the arguments suggest that both speed and intelligence might be pushed far beyond human capacity in a relatively short time. This process might not qualify as a 'singularity' in the strict sense from mathematics and physics, but it would be similar enough that the name is not altogether inappropriate.

The term 'singularity' was introduced[4] by the science fiction writer Vernor Vinge in a 1983 opinion article. It was brought into wider circulation by Vinge's influential 1993 article "The Coming Technological Singularity" and by the inventor and futurist Ray Kurzweil's popular 2005 book *The Singularity is Near*. In practice, the term is used in a number of different ways. A loose sense refers to phenomena whereby ever-more-rapid technological change leads to unpredictable consequences.[5] A very strict sense refers to a point where speed and intelligence go to infinity, as in the hypothetical speed/intelligence explosion above. Perhaps the core sense of the term, though, is a moderate sense in which it refers to an intelligence explosion through the recursive mechanism set out by I. J. Good, whether or not this intelligence explosion goes along with a speed explosion or with divergence to infinity. I will always use the term 'singularity' in this core sense in what follows.

One might think that the singularity would be of great interest to academic philosophers, cognitive scientists, and artificial intelligence researchers. In practice, this has not been the case.[6] Good was an eminent academic, but his article was largely unappreciated at the time. The subsequent discussion of the singularity has largely taken place in nonacademic circles, including Internet forums, popular media and books, and workshops organized by the independent Singularity Institute. Perhaps the highly speculative flavor of the singularity idea has been responsible for academic resistance.

I think this resistance is a shame, as the singularity idea is clearly an important one. The argument for a singularity is one that we should take seriously. And the questions surrounding the singularity are of enormous practical and philosophical concern.

Practically: If there is a singularity, it will be one of the most important events in the history of the planet. An intelligence explosion has enormous potential benefits: a cure for all known diseases, an end to poverty, extraordinary scientific advances, and much more. It also has enormous potential dangers: an end to the human race, an arms race of warring machines, the power to destroy the planet. So if there is even a small chance that there will be a singularity, we would do well to think about what forms it might take and whether there is anything we can do to influence the outcomes in a positive direction.

Philosophically: The singularity raises many important philosophical questions. The basic argument for an intelligence explosion is philosophically interesting in itself, and forces us to think hard about the nature of intelligence and about the mental capacities of artificial machines. The potential consequences of an intelligence explosion force us to think hard about values and morality and about consciousness and personal identity. In effect, the singularity brings up some of the hardest traditional questions in philosophy and raises some new philosophical questions as well.

Furthermore, the philosophical and practical questions intersect. To determine whether there might be an intelligence explosion, we need to better understand what intelligence is and whether machines might have it. To determine whether an intelligence explosion will be a good or a bad thing, we need to think about the relationship between intelligence and value. To determine whether we can play a significant role in a post-singularity world, we need to know whether human identity can survive the enhancing of our cognitive systems, perhaps through uploading onto new technology. These are life-or-death questions that may confront us in coming decades or centuries. To have any hope of answering them, we need to think clearly about the philosophical issues.

In what follows, I address some of these philosophical and practical questions. I start with the argument for a singularity: is there good reason to believe that there will be an intelligence explosion? Next, I consider how to negotiate the singularity: if it is possible that there will be a singularity, how can we maximize the chances of a good outcome? Finally, I consider the place of humans in a post-singularity world, with special attention to questions about uploading: can an uploaded human be conscious, and will uploading preserve personal identity?

My discussion will necessarily be speculative, but I think it is possible to reason about speculative outcomes with at least a modicum of rigor. For example, by formalizing arguments for a speculative thesis with premises and conclusions, one can see just what opponents need to be deny in order to deny the thesis, and one can then assess the costs of doing so. I will not try to give knockdown arguments in this paper, and I will not try to give final and definitive answers to the questions above, but I hope to encourage others to think about these issues further.[7]

2. The Argument for a Singularity

To analyze the argument for a singularity in a more rigorous form, it is helpful to introduce some terminology. Let us say that AI is artificial intelligence of human level or greater (that is, at least as intelligent as an average human). Let us say that AI+ is artificial intelligence of greater than human level (that is, more intelligent than the most intelligent human). Let us say that AI++ (or superintelligence) is AI of far greater than human level (say, at least as far beyond the most intelligent human as the most intelligent human is beyond a mouse).[8] Then we can put the argument for an intelligence explosion as follows:

1. There will be AI+.
2. If there is AI+, there will be AI++.

---

3. There will be AI++.

Here, premise 1 needs independent support (on which more soon), but is often taken to be plausible. Premise 2 is the key claim of the intelligence explosion, and is supported by Good's reasoning set out above. The conclusion says that there will be superintelligence.

The argument depends on the assumptions that there is such a thing as intelligence and that it can be compared between systems: otherwise the notion of an AI+ and an AI++ does not even make sense. Of course these assumptions might be questioned. Someone might hold that there is no single property that deserves

to be called 'intelligence,' or that the relevant properties cannot be measured and compared. For now, however, I will proceed with under the simplifying assumption that there is an intelligence measure that assigns an intelligence value to arbitrary systems. Later I will consider the question of how one might formulate the argument without this assumption. I will also assume that intelligence and speed are conceptually independent, so that increases in speed with no other relevant changes do not count as increases in intelligence.

We can refine the argument a little by breaking the support for premise 1 into two steps. We can also add qualifications about timeframe and about potential defeaters for the singularity.

1. There will be AI (before long, absent defeaters).
2. If there is AI, there will be AI+ (soon after, absent defeaters).
3. If there is AI+, there will be AI++ (soon after, absent defeaters).

---

4. There will be AI++ (before too long, absent defeaters).

Precise values for the timeframe variables are not too important. But we might stipulate that 'before long' means 'within centuries.' This estimate is conservative compared to those of many advocates of the singularity, who suggest decades rather than centuries. For example, Good 1965 predicts an ultraintelligent machine by 2000, Vinge 1993 predicts greater-than-human intelligence between 2005 and 2030, Yudkowsky 1996 predicts a singularity by 2021, and Kurzweil 2005 predicts human-level artificial intelligence by 2030.

Some of these estimates (e.g. Yudkowsky's) rely on extrapolating hardware trends.[9] My own view is that the history of artificial intelligence suggests that the biggest bottleneck on the path to AI is software, not hardware: we have to find the right algorithms, and no-one has come close to finding them yet. So I think that hardware extrapolation is not a good guide here. Other estimates (e.g. Kurzweil's) rely on estimates for when we will be able to artificially emulate an entire human brain. My sense is that most neuroscientists think these estimates are overoptimistic. Speaking for myself, I would be surprised if there were human-level AI within the next three decades. Nevertheless, my credence that there will be human-level AI before 2100 is somewhere over one-half. In any case, I think the move from decades to centuries renders the prediction conservative rather than radical, while still keeping the timeframe close enough to the present for the conclusion to be interesting.

By contrast, we might stipulate that 'soon after' means 'within decades.' Given the way that computer technology always advances, it is natural enough to think that once there is AI, AI+ will be just around the corner. And the argument for the intelligence explosion suggests a rapid step from AI+ to AI++ soon after that. I think it would not be unreasonable to suggest 'within years' here (and some would suggest 'within days' or even sooner for the second step), but as before 'within decades' is conservative while still being interesting. As for 'before too long,' we can stipulate that this is the sum of a 'before long' and two 'soon afters.' For present purposes, that is close enough to 'within centuries,' understood somewhat more loosely than the usage in the first premise to allow an extra century or so.

As for defeaters: I will stipulate that these are anything that prevents intelligent systems (human or artificial) from manifesting their capacities to create intelligent systems. Potential defeaters include disasters, disinclination, and active prevention.[10] For example, a nuclear war might set back our technological capacity enormously, or we (or our successors) might decide that a singularity would be a bad thing and prevent research that could bring it about. I do not think considerations internal to artificial intelligence can exclude these possibilities, although we might argue on other grounds about how likely they are. In any case, the notion of a defeater is still highly constrained (importantly, a defeater is *not* defined as anything that would prevent a singularity, which would make the conclusion near-trivial), and the conclusion that absent defeaters there will be superintelligence is strong enough to be interesting.

We can think of the three premises as an *equivalence* premise (there will be AI at least equivalent to our own intelligence), an *extension* premise (AI will soon be extended to AI+), and an *amplification* premise (AI+ will soon be greatly amplified to AI++). Why believe the premises? I will take them in order.

Premise 1: There will be AI (before long, absent defeaters).

One argument for the first premise is the *emulation argument*, based on the possibility of brain emulation. Here (following the usage of Sandberg and Bostrom 2008), emulation can be understood as close simulation: in this case,

simulation of internal processes in enough detail to replicate approximate patterns of behavior.

(i) The human brain is a machine.
(ii) We will have the capacity to emulate this machine (before long).
(iii) If we emulate this machine, there will be AI.

(iv) Absent defeaters, there will be AI (before long).

The first premise is suggested by what we know of biology (and indeed by what we know of physics). Every organ of the body appears to be a machine: that is, a complex system comprised of law-governed parts interacting in a law-governed way. The brain is no exception. The second premise follows from the claims that microphysical processes can be simulated arbitrarily closely and that any machine can be emulated by simulating microphysical processes arbitrarily closely. It is also suggested by the progress of science and technology more generally: we are gradually increasing our understanding of biological machines and increasing our capacity to simulate them, and there do not seem to be limits to progress here. The third premise follows from the definitional claim that if we emulate the brain this will replicate approximate patterns of human behaviour, along with the claim that such replication will result in AI. The conclusion follows from the premises along with the definitional claim that absent defeaters, systems will manifest their relevant capacities.

One might resist the argument in various ways. One could argue that the brain is more than a machine; one could argue that we will never have the capacity to emulate it; and one could argue that emulating it need not produce AI. Various existing forms of resistance to AI take each of these forms. For example, J. R. Lucas 1961 has argued that for reasons tied to Gödel's theorem, humans are more sophisticated than any machine. Hubert Dreyfus 1972 and Roger Penrose 1994 have argued that human cognitive activity can never be emulated by any computational machine. John Searle 1980 and Ned Block 1981 have argued that even if we can emulate the human brain, it does not follow that the emulation itself has a mind or is intelligent.

I have argued elsewhere that all of these objections fail.[11] But for present purposes, we can set many of them to one side. To reply to the Lucas, Penrose, and Dreyfus objections, we can note that nothing in the singularity idea requires that an AI be a *classical* computational system or even that it be a computational system at all. For example, Penrose (like Lucas) holds that the brain is not an algorithmic system in the ordinary sense, but he allows that it is a mechanical system that relies on certain nonalgorithmic quantum processes. Dreyfus holds that the brain is not a rule-following symbolic system, but he allows that it may nevertheless be a mechanical system that relies on subsymbolic processes (for example, connectionist processes). If so, then these arguments give us no reason to deny that we can build artificial systems that exploit the relevant nonalgorithmic quantum processes, or the relevant subsymbolic processes, and that thereby allow us to simulate the human brain.

As for the Searle and Block objections, these rely on the thesis that even if a system duplicates our behavior, it might be missing important 'internal' aspects of mentality: consciousness, understanding, intentionality, and so on. Later in the paper, I will advocate the view that if a system in our world duplicates not only our outputs but our internal computational structure, then it will duplicate the important internal aspects of mentality too. For present purposes, though, we can set aside these objections by stipulating that for the purposes of the argument, intelligence is to be measured wholly in terms of behavior and behavioral dispositions, where behavior is construed operationally in terms of the physical outputs that a system produces. The conclusion that there will be AI++ in this sense is still strong enough to be interesting. If there are systems that produce apparently superintelligent outputs, then whether or not these systems are truly conscious or intelligent, they will have a transformative impact on the rest of the world.

Perhaps the most important remaining form of resistance is the claim that the brain is not a mechanical system at all, or at least that nonmechanical processes play a role in its functioning that cannot be emulated. This view is most naturally combined with a sort of Cartesian dualism holding that some aspects of mentality (such as consciousness) are nonphysical and nevertheless play a substantial role in affecting brain processes and behavior. If there are nonphysical processes like this, it might be that they could nevertheless be emulated or artificially created, but this is not obvious. If these processes cannot be emulated or artificially created, then it may be that human-level AI is impossible.

Although I am sympathetic with some forms of dualism about consciousness, I do not think

that there is much evidence for the strong form of Cartesian dualism that this objection requires. The weight of evidence to date suggests that the brain is mechanical, and I think that even if consciousness plays a causal role in generating behavior, there is not much reason to think that its role is not emulable. But while we know as little as we do about the brain and about consciousness, I do not think the matter can be regarded as entirely settled. So this form of resistance should at least be registered.

Another argument for premise 1 is the *evolutionary argument*, which runs as follows.

(i) Evolution produced human-level intelligence.
(ii) If evolution produced human-level intelligence, then we can produce AI (before long).

(iii) Absent defeaters, there will be AI (before long).

Here, the thought is that since evolution produced human-level intelligence, this sort of intelligence is not entirely unattainable. Furthermore, evolution operates without requiring any antecedent intelligence or forethought. If evolution can produce something in this unintelligent manner, then in principle humans should be able to produce it much faster, by using our intelligence.

Again, the argument can be resisted, perhaps by denying that evolution produced intelligence, or perhaps by arguing that evolution produced intelligence by means of processes that we cannot mechanically replicate. The latter line might be taken by holding that evolution needed the help of superintelligent intervention, or needed the aid of other nonmechanical processes along the way, or needed an enormously complex history that we could never artificially duplicate, or needed an enormous amount of luck. Still, I think the argument makes at least a prima facie case for its conclusion.

We can clarify the case against resistance of this sort by changing 'Evolution produced human-level intelligence' to 'Evolution produced human-level intelligence mechanically and nonmiraculously' in both premises of the argument. Then premise (ii) is all the more plausible. Premise (i) will now be denied by those who think evolution involved nonmechanical processes, supernatural intervention, or extraordinary amounts of luck. But the premise remains plausible, and the structure of the argument is clarified.

Of course these arguments do not tell us how AI will first be attained. They suggest at least two possibilities: brain emulation (simulating the brain neuron by neuron) and artificial evolution (evolving a population of AIs through variation and selection). There are other possibilities: direct programming (writing the program for an AI from scratch, perhaps complete with a database of world knowledge), for example, and machine learning (creating an initial system and a learning algorithm that on exposure to the right sort of environment leads to AI). Perhaps there are others still. I doubt that direct programming is likely to be the successful route, but I do not rule out any of the others.

It must be acknowledged that every path to AI has proved surprisingly difficult to date. The history of AI involves a long series of optimistic predictions by those who pioneer a method, followed by a period of disappointment and reassessment. This is true for a variety of methods involving direct programming, machine learning, and artificial evolution, for example. Many of the optimistic predictions were not obviously unreasonable at the time, so their failure should lead us to reassess our prior beliefs in significant ways. It is not obvious just what moral should be drawn: Alan Perlis has suggested 'A year spent in artificial intelligence is enough to make one believe in God.' So optimism here should be leavened with caution. Still, my own view is that the balance of considerations still distinctly favors the view that AI will eventually be possible.

Premise 2: If there is AI, then there will be AI+ (soon after, absent defeaters).

One case for the extension premise comes from advances in information technology. Whenever we come up with a computational product, that product is soon afterwards obsolete due to technological advances. We should expect the same to apply to AI. Soon after we have produced a human-level AI, we will produce an even more intelligent AI: an AI+.

We might put the argument as follows.

(i) If there is AI, AI will be produced by an extendible method.
(ii) If AI is produced by an extendible method, we will have the capacity to extend the method (soon after).
(iii) Extending the method that produces an AI will yield an AI+.

(iv) Absent defeaters, if there is AI, there will (soon after) be AI+.

Here, an extendible method is a method that can easily be improved, yielding more intelligent systems. Given this definition, premises (ii) and (iii) follow immediately. The only question is premise (i).

Not every method of creating human-level intelligence is an extendible method. For example, the currently standard method of creating human-level intelligence is biological reproduction. But biological reproduction is not obviously extendible. If we have better sex, for example, it does not follow that our babies will be geniuses. Perhaps biological reproduction will be extendible using future technologies such as genetic engineering, but in any case the conceptual point is clear.

Another method that is not obviously extendible is brain emulation. Beyond a certain point, it is not the case that if we simply emulate brains better, then we will produce more intelligent systems. So brain emulation on its own is not clearly a path to AI+. It may nevertheless be that brain emulation speeds up the path to AI+. For example, emulated brains running on faster hardware or in large clusters might create AI+ much faster than we could without them. We might also be able to modify emulated brains in significant ways to increase their intelligence. We might use brain simulations to greatly increase our understanding of the human brain and of cognitive processing in general, thereby leading to AI+. But brain emulation will not on its own suffice for AI+: if it plays a role, some other path to AI+ will be required to supplement it.

Other methods for creating AI do seem likely to be extendible, however. For example, if we produce an AI by direct programming, then it is likely that like almost every program that has yet been written, the program will be improvable in multiple respects, leading soon after to AI+. If we produce an AI by machine learning, it is likely that soon after we will be able to improve the learning algorithm and extend the learning process, leading to AI+. If we produce an AI by artificial evolution, it is likely that soon after we will be able to improve the evolutionary algorithm and extend the evolutionary process, leading to AI+.

To make the case for premise (i), it suffices to make the case that either AI will be produced directly by an extendible method, or that if it is produced by a nonextendible method, this method will itself lead soon after to an extendible method. My own view is that both claims are plausible. I think that if AI is possible at all (as the antecedent of this premise assumes), then it should be possible to produce AI through a learning or evolutionary process, for example. I also think that if AI is produced through a nonextendible method such as brain emulation, this method is likely to greatly assist us in the search for an extendible method, along the lines suggested above. So I think there is good reason to believe premise (i).

To resist the premise, an opponent might suggest that we lie at a limit point in intelligence space: perhaps we are as intelligent as a system could be, or perhaps we are at least at a local maximum in that there is no easy path from systems like us to more intelligent systems. An opponent might also suggest that although intelligence space is not limited in this way, there are limits on our capacity to create intelligence, and that as it happens those limits lie at just the point of creating human-level intelligence. I think that there is not a great deal of antecedent plausibility to these claims, but again, the possibility of this form of resistance should at least be registered.

There are also potential paths to greater-than-human intelligence that do not rely on first producing AI and then extending the method. One such path is brain enhancement. We might discover ways to enhance our brains so that the resulting systems are more intelligent than any systems to date. This might be done genetically, pharmacologically, surgically, or even educationally. It might be done through implantation of new computational mechanisms in the brain, either replacing or extending existing brain mechanisms. Or it might be done simply by embedding the brain in an ever more sophisticated environment, producing an 'extended mind' (Clark and Chalmers 1998) whose capacities far exceed that of an unextended brain.

It is not obvious that enhanced brains should count as AI or AI+. Some potential enhancements will result in a wholly biological system, perhaps with artificially enhanced biological parts (where to be biological is to be based on DNA, let us say). Others will result in a system with both biological and nonbiological parts (where we might use organic DNA-based composition as a rough and ready criterion for being biological). At least in the near-term, all such systems will count as human, so there is a sense in which they do not have greater-than-human intelligence. For present purposes, I will stipulate that the baseline for human intelligence is set at current human standards, and I will stipulate that at least the systems with nonbiological components to their cognitive systems (brain

implants and technologically extended minds, for example) count as artificial. So intelligent enough systems of this sort will count as AI+.

Like other AI+ systems, enhanced brains suggest a potential intelligence explosion. An enhanced system may find further methods of enhancement that go beyond what we can find, leading to a series of ever-more-intelligent systems. Insofar as enhanced brains always rely on a biological core, however, there may be limitations. There are likely to be speed limitations on biological processing, and there may well be cognitive limitations imposed by brain architecture in addition. So beyond a certain point, we might expect non-brain-based systems to be faster and more intelligent than brain-based systems. Because of this, I suspect that brain enhancement that preserves a biological core is likely to be at best a first stage in an intelligence explosion. At some point, either the brain will be 'enhanced' in a way that dispenses with the biological core altogether, or wholly new systems will be designed. For this reason I will usually concentrate on non-biological systems in what follows. Still, brain enhancements raise many of the same issues and may well play an important role.

Premise 3: If there is AI+, there will be AI++ (soon after, absent defeaters).

The case for the amplification premise is essentially the argument from I. J. Good given above. We might lay it out as follows. Suppose there exists an AI+. Let us stipulate that AI_1 is the first AI+, and that AI_0 is its (human or artificial) creator. (If there is no sharp borderline between non-AI+ and AI+ systems, we can let AI_1 be any AI+ that is more intelligent than its creator.) Let us stipulate that δ is the difference in intelligence between AI_1 and AI_0, and that one system is significantly more intelligent than another if there is a difference of at least δ between them. Let us stipulate that for $n > 1$, an AI_{n+1} is an AI that is created by an AI_n and is significantly more intelligent than its creator.

(i) If there exists AI+, then there exists an AI_1.
(ii) For all $n > 0$, if an AI_n exists, then absent defeaters, there will be an AI_{n+1}.
(iii) If for all n there exists an AI_n, there will be AI++.

(iv) If there is AI+, then absent defeaters, there will be AI++.

Here premise (i) is true by definition. Premise (ii) follows from three claims: (a) the definitional claim that if AI_n exists, it is created by AI_{n-1} and is more intelligent than AI_{n-1}, (b) the definitional claim that if AI_n exists, then absent defeaters it will manifest its capacities to create intelligent systems, and (c) the substantive claim that if AI_n is significantly more intelligent than AI_{n-1}, it has the capacity to create a system significantly more intelligent than any that AI_{n-1} can create. Premise (iii) follows from the claim that if there is a sequence of AI systems each of which is significantly more intelligent than the last, there will eventually be superintelligence. The conclusion follows by logic and mathematical induction from the premises.

The conclusion as stated here omits the temporal claim 'soon after.' One can make the case for the temporal claim by invoking the ancillary premise that AI+ systems will be running on hardware much faster than our own, so that steps from AI+ onward are likely to be much faster than the step from humans to AI+.

There is room in logical space to resist the argument. For a start, one can note that the soundness of the argument depends on the intelligence measure used: if there is an intelligence measure for which the argument succeeds, there will almost certainly be a rescaled intelligence measure (perhaps a logarithmic measure) for which it fails. So for the argument to be interesting, we need to restrict it to intelligence measures that accord sufficiently well with intuitive intelligence measures that the conclusion captures the intuitive claim that there will be AI of far greater than human intelligence.

Relatedly, one could resist premise (iii) by holding that an arbitrary number of increases in intelligence by δ need not add up to the difference between AI+ and AI++. If we stipulate that δ is a ratio of intelligences, and that AI++ requires a certain fixed multiple of human intelligence (100 times, say), then resistance of this sort will be excluded. Of course for the conclusion to be interesting, then as in the previous paragraph, the intelligence measure must be such that this fixed multiple suffices for something reasonably counted as superintelligence.

The most crucial assumption in the argument lies in premise (ii) and the supporting claim (c). We might call this assumption a *proportionality thesis*: it holds that increases in intelligence (or increases of a certain sort) always lead to proportionate increases in the capacity to design intelligent systems. Perhaps the most promising way for an opponent to resist is to suggest that this thesis may fail. It might fail because here

are upper limits in intelligence space, as with resistance to the last premise. It might fail because there are points of diminishing returns: perhaps beyond a certain point, a 10% increase in intelligence yields only a 5% increase at the next generation, which yields only a 2.5% increase at the next generation, and so on. It might fail because intelligence does not correlate well with design capacity: systems that are more intelligent need not be better designers. I will return to resistance of these sorts in section 4, under 'structural obstacles.'

One might reasonably doubt that the proportionality thesis will hold across all possible systems and all the way to infinity. To handle such an objection, one can restrict premise (ii) to AI systems in a certain class. We just need some property ϕ such that an AI_n with ϕ can always produce an AI_{n+1} with ϕ, and such that we can produce an AI+ with ϕ. One can also restrict the proportionality thesis to a specific value of δ (rather than all possible values) and one can restrict n to a relatively small range $n < k$ (where $k = 100$, say) as long as k increases of δ suffices for superintelligence.

It is worth noting that in principle the recursive path to AI++ need not start at the human level. If we had a system whose overall intelligence were far lower than human level but which nevertheless had the capacity to improve itself or to design further systems, resulting in a system of significantly higher intelligence (and so on recursively), then the same mechanism as above would lead eventually to AI, AI+, and AI++. So in principle the path to AI++ requires only that we create a certain sort of self-improving system, and does not require that we directly create AI or AI+. In practice, the clearest case of a system with the capacity to amplify intelligence in this way is the human case (via the creation of AI+), and it is not obvious that there will be less intelligent systems with this capacity.[12] But the alternative hypothesis here should at least be noted.

3. The Intelligence Explosion Without Intelligence

The arguments so far have depended on an uncritical acceptance of the assumption that there is such a thing as intelligence and that it can be measured. Many researchers on intelligence accept these assumptions. In particular, it is widely held that there is such a thing as 'general intelligence,' often labeled g, that lies at the core of cognitive ability and that correlates with many different cognitive capacities.[13]

Still, many others questions these assumptions. Opponents hold that there is no such thing as intelligence, or at least that there is no single thing. On this view, there are many different ways of evaluating cognitive agents, no one of which deserves the canonical status of 'intelligence.' One might also hold that even if there is a canonical notion of intelligence that applies within the human sphere, it is far from clear that this notion can be extended to arbitrary non-human systems, including artificial systems. Or one might hold that the correlations between general intelligence and other cognitive capacities that hold within humans need not hold across arbitrary non-human systems. So it would be good to be able to formulate the key theses and arguments without assuming the notion of intelligence.

I think that this can be done. We can rely instead on the general notion of a cognitive capacity: some specific capacity that can be compared between systems. All we need for the purpose of the argument is (i) a self-amplifying cognitive capacity G: a capacity such that increases in that capacity go along with proportionate (or greater) increases in the ability to create systems with that capacity, (ii) the thesis that we can create systems whose capacity G is greater than our own, and (iii) a correlated cognitive capacity H that we care about, such that certain small increases in H can always be produced by large enough increases in G. Given these assumptions, it follows that absent defeaters, G will explode, and H will explode with it. (A formal analysis that makes the assumptions and the argument more precise follows at the end of the section.)

In the original argument, intelligence played the role of both G and H. But there are various plausible candidates for G and H that do not appeal to intelligence. For example, G might be a measure of programming ability, and H a measure of some specific reasoning ability. Here it is not unreasonable to hold that we can create systems with greater programming ability than our own, and that systems with greater programming ability will be able to create systems with greater programming ability in turn. It is also not unreasonable to hold that programming ability will correlate with increases in various specific reasoning abilities. If so, we should expect that absent defeaters, the reasoning abilities in question will explode.

This analysis brings out the importance of correlations between capacities in thinking about the singularity. In practice, we care about the singularity because we care about potential explosions in various specific capacities: the capacity to do science, to do philosophy, to create weapons, to take over the world, to bring about world peace, to be happy. Many or most of these capacities are not themselves self-amplifying, so we can expect an explosion in these capacities only to the extent that they correlate with other self-amplifying capacities. And for any given capacity, it is a substantive question whether they are correlated with self-amplifying capacity in this way. Perhaps the thesis is prima facie more plausible for the capacity to do science than for the capacity to be happy, but the questions are nontrivial.

The point applies equally to the intelligence analysis, which relies for its interest on the idea that intelligence correlates with various specific capacities. Even granted the notion of intelligence, the question of just what it correlates with is nontrivial. Depending on how intelligence is measured, we might expect it to correlate well with some capacities (perhaps a capacity to calculate) and to correlate less well with other capacities (perhaps a capacity for wisdom). It is also far from trivial that intelligence measures that correlate well with certain cognitive capacities within humans will also correlate with those capacities in artificial systems.

Still, two observations help with these worries. The first is that the correlations need not hold across all systems or even across all systems that we might create. There need only be some *type* of system such that the correlations hold across all systems of that type. If such a type exists (a subset of architectures, say), then recursive creation of systems of this type would lead to explosion. The second is that the self-amplifying capacity G need not correlate directly with the cognitive capacity H, but need only correlate with H', the capacity to create systems with H. While it is not especially plausible that design capacity will correlate with happiness, for example, it is somewhat more plausible that design capacity will correlate with the capacity to create happy systems. If so, then the possibility is left open that as design capacity explodes, happiness will explode along with it, either in the main line of descent or in a line of offshoots, at least if the designers choose to manifest their capacity to create happy systems.

A simple formal analysis follows (the remainder of this section can be skipped by those uninterested in formal details). Let us say that a parameter is a function from cognitive systems to positive real numbers. A parameter G *measures* a capacity C iff for all cognitive systems a and b, $G(a) > G(b)$ if a has a greater capacity C than b (one might also require that degrees of G correspond to degrees of C in some formal or intuitive sense). A parameter G *strictly tracks* a parameter H in ϕ-systems (where ϕ is some property or class of systems) if whenever a and b are ϕ-systems and $G(a) > G(b)$, then $H(a)/H(b) \geq G(a)/G(b)$. A parameter G *loosely tracks* a parameter H in ϕ-systems iff for all y there exists x such that (nonvacuously) for all ϕ-systems a, if $G(a) > x$, then $H(a) > y$. A parameter G strictly/loosely tracks a capacity C in ϕ-systems if it strictly/loosely tracks a parameter that measures C in ϕ-systems. Here, strict tracking requires that increases in G always produce proportionate increases in H, while loose tracking requires only that some small increase in H can always be produced by a large enough increase in G.

For any parameter G, we can define a parameter G': this is a parameter that measures a system's capacity to create systems with G. More specifically, $G'(x)$ is the highest value of h such that x has the capacity to create a system y such that $G(y) = h$. We can then say that G is a self-amplifying parameter (relative to x) if $G'(x) > G(x)$ and if G strictly tracks G' in systems downstream from x. Here a system is downstream from x if it is created through a sequence of systems starting from x and with ever-increasing values of G. Finally, let us say that for a parameter G or a capacity H, G++ and H++ systems are systems with values of G and capacities H that far exceed human levels.

Now we simply need the following premises:

(i) G is a self-amplifying parameter (relative to us).

(ii) G loosely tracks cognitive capacity H (downstream from us).

(iii) Absent defeaters, there will be G++ and H++.

The first half of the conclusion follows from premise (i) alone. Let AI_0 be us. If G is a self-amplifying parameter relative to us, then we are capable of creating a system AI_1 such that $G(AI_1) > G(AI_0)$. Let $\delta = G(AI_1)/G(AI_0)$. Because G strictly tracks G', $G'(AI_1) \geq \delta G'(AI_0)$. So AI_1 is capable of creating a system AI_2 such that

$G(AI_2) \geq \delta G(AI_1)$. Likewise, for all n, AI_n is capable of creating AI_{n+1} such that $G(AI_{n+1}) \geq \delta G(AI_n)$. It follows that absent defeaters, arbitrarily high values of G will be produced. The second half of the conclusion immediately follows from (ii) and the first half of the conclusion. Any value of H can be produced by a high enough value of G, so it follows that arbitrarily high values for H will be produced.

The assumptions can be weakened in various ways. As noted earlier, it suffices for G to loosely track not H but H', where H' measures the capacity to create systems with H. Furthermore, the tracking relations between G and G', and between G and H or H', need not hold in all systems downstream from us: it suffices that there is a type ϕ such that in ϕ-systems downstream from us, G strictly tracks $G'(\phi)$ (the ability to create a ϕ-system with G) and loosely tracks H or H'. We need not require that G is strictly self-amplifying: it suffices for G and H (or G and H') to be jointly self-amplifying in that high values of both G and H lead to significantly higher values of each. We also need not require that the parameters are self-amplifying forever. It suffices that G is self-amplifying over however many generations are required for $G++$ (if $G++$ requires a 100-fold increase in G, then $\log_\delta 100$ generations will suffice) and for $H++$ (if $H++$ requires a 100-fold increase in H and the loose tracking relation entails that this will be produced by an increase in G of 1000, then $\log_\delta 1000$ generations will suffice). Other weakenings are also possible.

4. Obstacles to the Singularity

On the current analysis, an intelligence explosion results from a self-amplifying cognitive capacity (premise (i) above), correlations between that capacity and other important cognitive capacities (premise (ii) above), and manifestation of those capacities (conclusion). More pithily: self-amplification plus correlation plus manifestation = singularity.

This analysis brings out a number of potential obstacles to the singularity: that is, ways that there might fail to be a singularity. There might fail to be interesting self-amplifying capacities. There might fail to be interesting correlated capacities. Or there might be defeaters, so that these capacities are not manifested. We might call these *structural obstacles*, *correlation obstacles*, and *manifestation obstacles* respectively.

I do not think that there are knockdown arguments against any of these three sorts of obstacles. I am inclined to think that manifestation obstacles are the most serious obstacle, however. I will briefly discuss obstacles of all three sorts in what follows.

Structural obstacles: There are three overlapping ways in which there might fail to be relevant self-amplifying capacities, which we can illustrate by focusing on the case of intelligence. *Limits in intelligence space*: we are at or near an upper limit in intelligence space. *Failure of takeoff*: although there are higher points in intelligence space, human intelligence is not at a takeoff point where we can create systems more intelligent than ourselves. *Diminishing returns*: although we can create systems more intelligent than ourselves, increases in intelligence diminish from there. So a 10% increase might lead to a 5% increase, a 2.5% increase, and so on, or even to no increase at all after a certain point.

Regarding limits in intelligence space: While the laws of physics and the principles of computation may impose limits on the sort of intelligence that is possible in our world, there is little reason to think that human cognition is close to approaching those limits. More generally, it would be surprising if evolution happened to have recently hit or come close to an upper bound in intelligence space.

Regarding failure of takeoff: I think that the prima facie arguments earlier for AI and AI+ suggest that we are at a takeoff point for various capacities such as the ability to program. There is prima facie reason to think that we have the capacity to emulate physical systems such as brains. And there is prima facie reason to think that we have the capacity to improve on those systems.

Regarding diminishing returns: These pose perhaps the most serious structural obstacle. Still, I think there is some plausibility in proportionality theses, at least given an intuitive intelligence measure. If anything, 10% increases in intelligence-related capacities are likely to lead all sorts of intellectual breakthroughs, leading to next-generation increases in intelligence that are significantly greater than 10%. Even among humans, relatively small differences in design capacities (say, the difference between Turing and an average human) seem to lead to large differences in the systems that are designed (say, the difference between a computer and nothing of importance). And even if there are diminishing returns, a limited increase in intelligence combined with a large increase in speed

will produce at least some of the effects of an intelligence explosion.

One might worry that a 'hill-climbing' process that starts from the human cognitive system may run into a local maximum from which one cannot progress further by gradual steps. I think that this possibility is made less likely by the enormous dimensionality of intelligence space and by the enormous number of paths that are possible. In addition, the design of AI is not limited to hill-climbing: there is also 'hill-leaping,' where one sees a favorable area of intelligence space some distance away and leaps to it. Perhaps there are some areas of intelligence space (akin to inaccessible cardinals in set theory?) that one simply cannot get to by hill-climbing and hill-leaping, but I think that there is good reason to think that these processes at least can get us far beyond ordinary human capacities.

Correlation obstacles. It may be that while there is one or more self-amplifying cognitive capacity *G*, this does not correlate with any or many capacities that are of interest to us. For example, perhaps a self-amplifying increase in programming ability will not go along with increases in other interesting abilities, such as an ability to solve scientific problems or social problems, an ability to wage warfare or make peace, and so on.

I have discussed issues regarding correlation in the previous section. I think that the extent to which we can expect various cognitive capacities to correlate with each other is a substantive open question. Still, even if self-amplifying capacities such as design capacities correlate only weakly with many cognitive capacities, they will plausibly correlate more strongly with the capacity to create systems with these capacities. It remains a substantive question just how much correlation one can expect, but I suspect that there will be enough correlating capacities to ensure that if there is an explosion, it will be an interesting one.

Manifestation obstacles. Although there is a self-amplifying cognitive capacity *G*, either we or our successors might not manifest our capacity to create systems with higher values of *G* (or with higher values of a cognitive correlated capacity *H*). Here we can divide the defeaters into *motivational defeaters* in which an absence of motivation or a contrary motivation prevents capacities from being manifested, and *situational defeaters*, in which other unfavorable circumstances prevent capacities from being manifested. Defeaters of each sort could arise on the path to AI, on the path from AI to AI+, or on the path from AI+ to AI++.

Situational defeaters include disasters and resource limitations. Regarding disasters, I certainly cannot exclude the possibility that global warfare or a nanotechnological accident ('gray goo') will stop technological progress entirely before AI or AI+ is reached. I also cannot exclude the possibility that artificial systems will themselves bring about disasters of this sort. Regarding resource limitations, it is worth noting that most feedback loops in nature run out of steam because of limitations in resources such as energy, and the same is possible here. Still, it is likely that foreseeable energy resources will suffice for many generations of AI+, and AI+ systems are likely to develop further ways of exploiting energy resources. Something similar applies to financial resources and other social resources.

Motivational defeaters include disinclination and active prevention. It is possible that as the event draws closer, most humans will be disinclined to create AI or AI+. It is entirely possible that there will be active prevention of the development of AI or AI+ (perhaps by legal, financial, and military means), although it is not obvious that such prevention could be successful indefinitely.[14] And it is certainly possible that AI+ systems will be disinclined to create their successors, perhaps because we design them to be so disinclined, or perhaps because they will be intelligent enough to realize that creating successors is not in their interests. Furthermore, it may be that AI+ systems will have the capacity to prevent such progress from happening.

A singularity proponent might respond that all that is needed to overcome motivational defeaters is the creation of a single AI+ that greatly values the creation of greater AI+ in turn, and a singularity will then be inevitable. If such a system is the first AI+ to be created, this conclusion may well be correct. But as long as this AI+ is not created first, then it may be subject to controls from other AI+, and the path to AI++ may be blocked. The issues here turn on difficult questions about the motivations and capacities of future systems, and answers to these questions are difficult to predict.

In any case, the current analysis makes clearer the burdens on both proponents and opponents of the thesis that there will be an intelligence explosion. Opponents need to make clear where they think the case for the thesis fails: structural obstacles (and if so which),

correlation obstacles, situational defeaters, motivational defeaters. Likewise, proponents need to make the case that there will be no such obstacles or defeaters.

Speaking for myself, I think that while structural and correlational obstacles (especially the proportionality thesis) raise nontrivial issues, there is at least a prima facie case that *absent defeaters*, a number of interesting cognitive capacities will explode. I think the most likely defeaters are motivational. But I think that it is far from obvious that there will be defeaters. So I think that the singularity hypothesis is one that we should take very seriously.

BIBLIOGRAPHY

Block, N., "Psychologism and behaviorism," *Philosophical Review* 90 (1981): pp. 5–43.

Bostrom, N., "How long before superintelligence?" *International Journal of Future Studies* 2 (1998), http://www.nickbostrom.com/superintelligence.html.

_____ "Ethical issues in advanced artificial intelligence," in *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, vol. 2, I. Smit, ed., (International Institute of Advanced Studies in Systems Research and Cybernetics, 2003).

_____ "Quantity of experience: Brain-duplication and degrees of consciousness," *Minds and Machines* 16 (2006): pp. 185–200.

Campbell, J. W., "The last evolution," *Amazing Stories* 1932).

Chalmers, D. J., :How Cartesian dualism might have been true," 1990.

_____ "Minds, machines, and mathematics," *Psyche* 2 (1995): pp. 11–20.

_____ *The Conscious Mind* (New York: Oxford University Press, 1996).

_____ "Consciousness and its place in nature," in *Blackwell Guide to the Philosophy of Mind*, S. Stich and F. Warfield, eds., (Oxford: Blackwell, 2003).

_____ "The Matrix as metaphysics," in *Philosophers Explore the Matrix* C. Grau, ed., (New York: Oxford University Press, 2005).

_____ "Perception and the fall from Eden," in *Perceptual Experience*, T. Gendler and J. Hawthorne, eds., (New York: Oxford University Press, 2006.

Clark, A., and Chalmers, D., "The extended mind," *Analysis* 58 (1998): pp. 7–19.

Dainton, B. *The Phenomenal Self* (New York: Oxford University Press, 2008).

Dreyfus, H. *What Computers Can't Do* (Cambridge, MA: MIT Press, 1972).

Egan, G. *Permutation City* (London: Orion Books/ Millenium, 1994).

Floridi, L., and Sanders, J. W., "On the morality of artificial agents," *Minds and Machines* 14 (2004): pp. 349–79.

Flynn, J. R. *What is Intelligence?* (Cambridge, UK: Cambridge University Press, 2007).

Good, I. J., "Speculations concerning the first ultraintelligent machine," in *Advances in Computers*, vol. 6, F. Alt and M. Rubinoff, eds., (Academic Press, 1965).

Hanson, R., "If uploads come first: The crack of a future dawn," *Extropy* 6 (1994): p. 2, http://hanson.gmu.edu/uploads.html.

_____ "Economics of the singularity," *IEEE Spectrum* (June 2008): pp. 37–43.

_____ "Prefer law to values," *Overcoming Bias*, October 10, 2009, http://www.overcomingbias.com/2009/10/prefer-law-to-values.html.

Hofstadter, D. R., "Moores law, artificial evolution, and the fate of humanity," in *Perspectives on Adaptation in Natural and Artificial Systems*, L. Booker, S. Forrest, M. Mitchell, and R. Riolo, eds., (New York: Oxford University Press, 2005).

Joy, W., "Why the future doesn't need us," *Wired* 8, no. 04 (July 2000).

Kurzweil, R. *The Singularity is Near* (New York: Penguin Books, 2005).

Legg, S. *Machine Superintelligence* (2008).

Lucas, J. R., "Minds, machines, and Gödel," *Philosophy* 36 (1961): pp. 112–27.

Moravec, H. *Mind Children: The Future of Robot and Human Intelligence* (Cambridge, MA: Harvard University Press, 1988).

_____ *Robots: Mere Machine to Transcendent Mind* (New York: Oxford University Press, 1998).

Omohundro, S., "The nature of self-improving artificial intelligence," 2007, http://steveomohundro.com/scientific-contributions.

_____ "The basic AI drives," in *Proceedings of the First AGI Conference*, Frontiers in Artificial Intelligence and Applications, vol. 171, P. Wang, B. Goertzel, and S. Franklin, eds., (IOS Press, 2008).

Parfit, D. A. *Reasons and Persons* (New York: Oxford University Press, 1984).

Penrose, R. *Shadows of the Mind* (Oxford University Press, 1994).

Sandberg, A., and Bostrom, N., "Whole brain emulation: A roadmap," Future for Humanity Institute, Oxford University, 2008.

Sawyer, R. *Calculating God*, (New York: Tor 2000).

_____ *Mindscan* (New York: Tor, 2005).

Schmidhuber, J., "Gödel machines: Self-referential universal problem solvers making provably optimal self-improvements," *arXiv* 2003, http://arxiv.org/abs/cs.LO/0309048.

Searle, J., "Minds, brains, and programs," *Behavioral and Brain Sciences* 3 (1980): pp. 417–57.

Shalizi, C., "*g*, a Statistical Myth," 2007, http://bactra.org/weblog/523.html.

Smart, J., Brief history of intellectual discussion of accelerating change," 1999–2008

Solomonoff, F., "1985. The time scale of artificial intelligence: Reflections on social effects. North-Holland Human Systems Management," 5:149–153. Elsevier.

Strout, J., *The mind uploading home page*, 2006, http://www.ibiblio.org/jstrout/uploading/.

Ulam, S., "John von Neumann 1903–1957," *Bulletin of the American Mathematical Society* 64, no, 3, part 2 (1958): pp. 1–49.

Unger, P. *Identity, Consciousness, and Value* (New York: Oxford University Press, 1990).

Vinge, V. "First word," *Omni* (January 1983): p. 10.

_____ "The coming technological singularity: How to survive in the post-human era," *Whole Earth Review* (winter 1993).

Wallach, W., and Allen, C. *Moral Machines: Teaching Robots Right from Wrong* (New York: Oxford University Press, 2009).

Yudkowsky, E., "Staring at the singularity," 1996.

_____ "The AI-box experiment," 2002, http://yudkowsky.net/singularity/aibox.

_____ "Three major singularity schools," 2007, http://yudkowsky.net/singularity/schools.

_____ "Artificial intelligence as a positive and negative factor in global risk," in *Global Catastrophic Risks*, N. Bostrom, ed., (New York: Oxford University Press, 2008).

NOTES

1. I first became interested in this cluster of ideas as a student, before first hearing explicitly of the 'singularity' in 1997. I was spurred to think further about these issues by an invitation to speak at the 2009 Singularity Summit in New York City. I thank many people at that event for discussion, as well as many at later talks and discussions at West Point, CUNY, New York University, Delhi, ANU, Tucson, Oxford, and UNSW. Thanks also to Doug Hofstadter, Marcus Hutter, Ole Koksvik, Drew McDermott, Carl Shulman, and Michael Vassar for comments on this paper.
2. Scenarios of this sort have antecedents in science fiction, perhaps most notably in John Campbell's 1932 short story "The Last Evolution."
3. Solomonoff also discusses the effects of what we might call the 'population explosion': a rapidly increasing population of artificial AI researchers.
4. As Vinge 1993 notes, Stanislaw Ulam 1958 describes a conversation with John von Neumann in which the term is used in a related way: 'One conversation centered on the ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue.'
5. A useful taxonomy of uses of 'singularity' is set out by Yudkowsky 2007. He distinguishes an 'accelerating change' school, associated with Kurzweil, an 'event horizon' school, associated with Vinge, and an 'intelligence explosion' school, associated with Good. Smart (1999–2008) gives a detailed history of associated ideas, focusing especially on accelerating change.
6. With some exceptions: discussions by academics include 1998, Bostrom 2003; Hanson 2008; Hofstadter 2005; and Moravec 1988, 1998. Hofstadter organized symposia on the prospect of superintelligent machines at Indiana University in 1999 and at Stanford University in 2000, and more recently, Bostrom's Future of Humanity Institute at the University of Oxford has organized a number of relevant activities.
7. The main themes in this article have been discussed many times before by others, especially in the non-academic circles mentioned earlier. My main aims in writing the article are to subject some of these themes (especially the claim that there will be an intelligence explosion and claims about uploading) to a philosophical analysis, with the aim of exploring and perhaps strengthening the foundations on which these ideas rest, and also to help bring these themes to the attention of philosophers and scientists.
8. Following common practice, I use 'AI' and relatives as a general term ('An AI exists'), an adjective ('An AI system exists'), and as a mass term ('AI exists').
9. Yudkowsky's web-based article is now marked 'obsolete,' and in later work he does not endorse the estimate or the argument from hardware trends. See Hofstadter 2005 for skepticism about the role of hardware extrapolation here and more generally for skepticism about timeframe estimates on the order of decades.
10. I take it that when someone has the capacity to do something, then if they are sufficiently motivated to do it and are in reasonably favorable circumstances, they will do it. So defeaters can be divided into *motivational defeaters*, involving insufficient motivation, and *situational defeaters*, involving unfavorable circumstances (such as a disaster). There is a blurry line between unfavorable circumstances that prevent a capacity from being manifested and those that entail that the capacity was never present in the first place—for example, resource limitations might be classed on either side of this line—but this will not matter much for our purposes.
11. For a general argument for strong artificial intelligence and a response to many different objections, see Chalmers 1996, chapter 9. For a response to Penrose and Lucas, see Chalmers 1995. For an in-depth discussion of the current prospects for whole brain emulation, see Sandberg and Bostrom 2008.
12. The 'Gödel machines' of Schmidhuber 2003 provide a theoretical example of self-improving systems at a level below AI, though they have not yet been implemented and there are large practical obstacles to using them as a path to AI. The process of evolution might count as an indirect example: less intelligent systems have the capacity to create more intelligent systems by reproduction, variation and natural selection. This version would then come to the same thing as an evolutionary path to AI and AI++. For present purposes I am construing 'creation' to involve a more direct mechanism than this.

13. Flynn 2007 gives an excellent overview of the debate over general intelligence and the reasons for believing in such a measure. Shalizi 2007 argues that *g* is a statistical artifact. Legg 2008 has a nice discussion of these issues in the context of machine superintelligence.
14. When I discussed these issues with cadets and staff at the West Point Military Academy, the question arose as to whether the US military or other branches of the government might attempt to prevent the creation of AI or AI+, due to the risks of an intelligence explosion. The consensus was that they would not, as such prevention would only increase the chances that AI or AI+ would first be created by a foreign power. One might even expect an AI arms race at some point, once the potential consequences of an intelligence explosion are registered. According to this reasoning, although AI+ would have risks from the standpoint of the US government, the risks of Chinese AI+ (say) would be far greater.

The Ethics of Artificial Intelligence

Nick Bostrom and Eliezer Yudkowsky

The possibility of creating thinking machines raises a host of ethical issues. These questions relate both to ensuring that such machines do not harm humans and other morally relevant beings, and to the moral status of the machines themselves. The first section discusses issues that may arise in the near future of AI. The second section outlines challenges for ensuring that AI operates safely as it approaches humans in its intelligence. The third section outlines how we might assess whether, and in what circumstances, AIs themselves have moral status. In the fourth section, we consider how AIs might differ from humans in certain basic respects relevant to our ethical assessment of them. The final section addresses the issues of creating AIs more intelligent than human, and ensuring that they use their advanced intelligence for good rather than ill.

Ethics in Machine Learning and Other Domain-Specific AI Algorithms

Imagine, in the near future, a bank using a machine learning algorithm to recommend mortgage applications for approval. A rejected applicant brings a lawsuit against the bank, alleging that the algorithm is discriminating racially against mortgage applicants. The bank replies that this is impossible, since the algorithm is deliberately blinded to the race of the applicants. Indeed, that was part of the bank's rationale for implementing the system. Even so, statistics show that the bank's approval rate for black applicants has been steadily dropping. Submitting ten apparently equally qualified genuine applicants (as determined by a separate panel of human judges) shows that the algorithm accepts white applicants and rejects black applicants. What could possibly be happening?

Finding an answer may not be easy. If the machine learning algorithm is based on a complicated neural network, or a genetic algorithm produced by directed evolution, then it may prove nearly impossible to understand why, or even how, the algorithm is judging applicants based on their race. On the other hand, a machine learner based on decision trees or Bayesian networks is much more transparent to programmer inspection (Hastie *et al.* 2001), which may enable an auditor to discover that the AI algorithm uses the address information of applicants who were born or previously resided in predominantly poverty-stricken areas.

AI algorithms play an increasingly large role in modern society, though usually not labeled 'AI.' The scenario described above might be transpiring even as we write. It will become increasingly important to develop AI algorithms that are not just powerful and scalable, but also

From (Keith Frankish and William M. Ramsey, eds.) *The Cambridge Handbook of Artificial Intelligence* (Cambridge University Press, 2014). Reprinted by permission of the publisher.

transparent to inspection—to name one of many socially important properties.

Some challenges of machine ethics are much like many other challenges involved in designing machines. Designing a robot arm to avoid crushing stray humans is no more morally fraught than designing a flame-retardant sofa. It involves new programming challenges, but no new ethical challenges. But when AI algorithms take on cognitive work with social dimensions—cognitive tasks previously performed by humans—the AI algorithm inherits the social requirements. It would surely be frustrating to find that no bank in the world will approve your seemingly excellent loan application, and nobody knows why, and nobody can find out even in principle. (Maybe you have a first name strongly associated with deadbeats? Who knows?)

Transparency is not the only desirable feature of AI. It is also important that AI algorithms taking over social functions be *predictable to those they govern.* To understand the importance of such predictability, consider an analogy. The legal principle of *stare decisis* binds judges to follow past precedent whenever possible. To an engineer, this preference for precedent may seem incomprehensible—why bind the future to the past, when technology is always improving? But one of the most important functions of the legal system is to be predictable, so that, e.g., contracts can be written knowing how they will be executed. The job of the legal system is not necessarily to optimize society, but to provide a predictable environment within which citizens can optimize their own lives.

It will also become increasingly important that AI algorithms be *robust against manipulation.* A machine vision system to scan airline luggage for bombs must be robust against human adversaries deliberately searching for exploitable flaws in the algorithm—for example, a shape that, placed next to a pistol in one's luggage, would neutralize recognition of it. Robustness against manipulation is an ordinary criterion in information security; nearly *the* criterion. But it is not a criterion that appears often in machine learning journals, which are currently more interested in, e.g., how an algorithm scales up on larger parallel systems.

Another important social criterion for dealing with organizations is being able to find the person responsible for getting something done. When an AI system fails at its assigned task, who takes the blame? The programmers? The end-users? Modern bureaucrats often take refuge in established procedures that distribute responsibility so widely that no one person can be identified to blame for the catastrophes that result (Howard 1994). The provably disinterested judgment of an expert system could turn out to be an even better refuge. Even if an AI system is designed with a user override, one must consider the career incentive of a bureaucrat who will be personally blamed if the override goes wrong, and who would much prefer to blame the AI for any difficult decision with a negative outcome.

Responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream with helpless frustration: all criteria that apply to humans performing social functions; all criteria that must be considered in an algorithm intended to replace human judgment of social functions; all criteria that may not appear in a journal of machine learning considering how an algorithm scales up to more computers. This list of criteria is by no means exhaustive, but it serves as a small sample of what an increasingly computerized society should be thinking about.

Artificial General Intelligence

There is nearly universal agreement among modern AI professionals that Artificial Intelligence falls short of human capabilities in some critical sense, even though AI algorithms have beaten humans in many specific domains such as chess. It has been suggested by some that as soon as AI researchers figure out how to do something, that capability ceases to be regarded as intelligent—chess was considered the epitome of intelligence until Deep Blue won the world championship from Kasparov—but even these researchers agree that something important is missing from modern AIs (e.g., Hofstadter 2006).

While this subfield of Artificial Intelligence is only just coalescing, 'Artificial General Intelligence' (hereafter, AGI) is the emerging term of art used to denote 'real' AI (see, e.g., the edited volume Goertzel and Pennachin 2006). As the name implies, the emerging consensus is that the missing characteristic is generality. Current AI algorithms with human-equivalent or -superior performance are characterized by a deliberately-programmed competence only in a single, restricted domain. Deep Blue became the world champion at chess, but it cannot even play checkers, let alone drive a car or

make a scientific discovery. Such modern AI algorithms resemble all biological life with the sole exception of *Homo sapiens*. A bee exhibits competence at building hives; a beaver exhibits competence at building dams; but a bee doesn't build dams, and a beaver can't learn to build a hive. A human, watching, can learn to do both; but this is a unique ability among biological lifeforms. It is debatable whether human intelligence is truly *general*—we are certainly better at some cognitive tasks than others (Hirschfeld and Gelman 1994)—but human intelligence is surely *significantly more generally applicable* than nonhominid intelligence.

It is relatively easy to envisage the sort of safety issues that may result from AI operating only within a specific domain. It is a qualitatively different class of problem to handle an AGI operating across many novel contexts that cannot be predicted in advance.

When human engineers build a nuclear reactor, they envision the specific events that could go on inside it—valves failing, computers failing, cores increasing in temperature—and engineer the reactor to render these events noncatastrophic. Or, on a more mundane level, building a toaster involves envisioning bread and envisioning the reaction of the bread to the toaster's heating element. The toaster itself does not know that its purpose is to make toast—the *purpose* of the toaster is represented within the designer's mind, but is not explicitly represented in computations inside the toaster—and so if you place cloth inside a toaster, it may catch fire, as the design executes in an unenvisioned context with an unenvisioned side effect.

Even task-specific AI algorithms throw us outside the toaster-paradigm, the domain of locally preprogrammed, specifically envisioned behavior. Consider Deep Blue, the chess algorithm that beat Garry Kasparov for the world championship of chess. Were it the case that machines can only do exactly as they are told, the programmers would have had to manually preprogram a database containing moves for every possible chess position that Deep Blue could encounter. But this was not an option for Deep Blue's programmers. First, the space of possible chess positions is unmanageably large. Second, if the programmers had manually input what *they* considered a good move in each possible situation, the resulting system would not have been able to make stronger chess moves than its creators. Since the programmers themselves were not world champions, such a system would not have been able to defeat Garry Kasparov.

In creating a superhuman chess player, the human programmers necessarily sacrificed their ability to predict Deep Blue's *local, specific* game behavior. Instead, Deep Blue's programmers had (justifiable) confidence that Deep Blue's chess moves would satisfy a *non-local* criterion of optimality: namely, that the moves would tend to steer the future of the game board into outcomes in the 'winning' region as defined by the chess rules. This prediction about distant consequences, though it proved accurate, did not allow the programmers to envision the *local* behavior of Deep Blue—its response to a specific attack on its king—because Deep Blue computed the nonlocal game map, the link between a move and its possible future consequences, more accurately than the programmers could (Yudkowsky 2006).

Modern humans do literally millions of things to feed themselves—to serve the final consequence of being fed. Few of these activities were 'envisioned by Nature' in the sense of being ancestral challenges to which we are directly adapted. But our adapted brain has grown powerful enough to be *significantly more generally applicable;* to let us foresee the consequences of millions of different actions across domains, and exert our preferences over final outcomes. Humans crossed space and put footprints on the Moon, even though none of our ancestors encountered a challenge analogous to vacuum. Compared to domain-specific AI, it is a qualitatively different problem to design a system that will operate safely across thousands of contexts; including contexts not specifically envisioned by either the designers or the users; including contexts that no human has yet encountered. Here there may be no *local* specification of good behavior—no simple specification over the behaviors themselves, any more than there exists a compact local description of all the ways that humans obtain their daily bread.

To build an AI that acts safely while acting in many domains, with many consequences, including problems the engineers never explicitly envisioned, one must specify good behavior in such terms as 'X such that the consequence of X is not harmful to humans.' This is non-local; it involves extrapolating the distant consequences of actions. Thus, this is only an effective specification—one that can be realized as a design property—if the system explicitly extrapolates the consequences of its behavior. A toaster cannot have this design property because

a toaster cannot foresee the consequences of toasting bread.

Imagine an engineer having to say, 'Well, I have no idea how this airplane I built will fly safely—indeed I have no idea how it will fly at all, whether it will flap its wings or inflate itself with helium or something else I haven't even imagined—but I assure you, the design is very, very safe.' This may seem like an unenviable position from the perspective of public relations, but it's hard to see what other guarantee of ethical behavior would be possible for a general intelligence operating on unforeseen problems, across domains, with preferences over distant consequences. Inspecting the cognitive design might verify that the mind was, indeed, searching for solutions that we would classify as ethical; but we couldn't predict which specific solution the mind would discover.

Respecting such a verification requires some way to distinguish trustworthy assurances (a procedure which will not say the AI is safe unless the AI really is safe) from pure hope and magical thinking ('I have no idea how the Philosopher's Stone will transmute lead to gold, but I assure you, it will!'). One should bear in mind that purely hopeful expectations have previously been a problem in AI research (McDermott 1976).

Verifiably constructing a trustworthy AGI will require different methods, and a different way of thinking, from inspecting power plant software for bugs—it will require an AGI that *thinks like* a human engineer concerned about ethics, not just a simple *product* of ethical engineering.

Thus the discipline of AI ethics, especially as applied to AGI, is likely to differ fundamentally from the ethical discipline of noncognitive technologies, in that:

- The local, specific behavior of the AI may not be predictable apart from its safety, even if the programmers do everything right;
- Verifying the safety of the system becomes a greater challenge because we must verify what the system is trying to do, rather than being able to verify the system's safe behavior in all operating contexts;
- Ethical cognition itself must be taken as a subject matter of engineering.

Machines with Moral Status

A different set of ethical issues arises when we contemplate the possibility that some future AI systems might be candidates for having moral status. Our dealings with beings possessed of moral status are not exclusively a matter of instrumental rationality: we also have moral reasons to treat them in certain ways, and to refrain from treating them in certain other ways. Francis Kamm has proposed the following definition of moral status, which will serve for our purposes:

> X has moral status = because X counts morally in its own right, it is permissible/impermissible to do things to it for its own sake. (Kamm 2007, chapter 7, paraphrase)

A rock has no moral status: we may crush it, pulverize it, or subject it to any treatment we like without any concern for the rock itself. A human person, on the other hand, must be treated not only as a means but also as an end. Exactly what it means to treat a person as an end is something about which different ethical theories disagree; but it certainly involves taking her legitimate interests into account—giving weight to her well-being—and it may also involve accepting strict moral side-constraints in our dealings with her, such as a prohibition against murdering her, stealing from her, or doing a variety of other things to her or her property without her consent. Moreover, it is because a human person counts in her own right, and for her sake, that it is impermissible to do to her these things. This can be expressed more concisely by saying that a human person has moral status.

Questions about moral status are important in some areas of practical ethics. For example, disputes about the moral permissibility of abortion often hinge on disagreements about the moral status of the embryo. Controversies about animal experimentation and the treatment of animals in the food industry involve questions about the moral status of different species of animal. And our obligations towards human beings with severe dementia, such as late-stage Alzheimer's patients, may also depend on questions of moral status.

It is widely agreed that current AI systems have no moral status. We may change, copy, terminate, delete, or use computer programs as we please; at least as far as the programs themselves are concerned. The moral constraints to which we are subject in our dealings with contemporary AI systems are all grounded in our responsibilities to other beings, such as our fellow humans, not in any duties to the systems themselves.

While it is fairly consensual that present-day AI systems lack moral status, it is unclear exactly what attributes ground moral status.

Two criteria are commonly proposed as being importantly linked to moral status, either separately or in combination: sentience and sapience (or personhood). These may be characterized roughly as follows:

> Sentience: the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer.
>
> Sapience: a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent.

One common view is that many animals have qualia and therefore have some moral status, but that only human beings have sapience, which gives them a higher moral status than non-human animals.[1] This view, of course, must confront the existence of borderline cases such as, on the one hand, human infants or human beings with severe mental retardation—sometimes unfortunately referred to as 'marginal humans'—which fail to satisfy the criteria for sapience; and, on the other hand, some non-human animals such as the great apes, which might possess at least some of the elements of sapience. Some deny that so-called 'marginal humans' have full moral status. Others propose additional ways in which an object could qualify as a bearer of moral status, such as by being a member of a kind that normally has sentience or sapience, or by standing in a suitable relation to some being that independently has moral status (cf. Mary Anne Warren 2000). For present purposes, however, we will focus on the criteria of sentience and sapience.

This picture of moral status suggests that an AI system will have some moral status if it has the capacity for qualia, such as an ability to feel pain. A sentient AI system, even if it lacks language and other higher cognitive faculties, is not like a stuffed toy animal or a wind-up doll; it is more like a living animal. It is wrong to inflict pain on a mouse, unless there are sufficiently strong morally overriding reasons to do so. The same would hold for any sentient AI system. If in addition to sentience, an AI system also has sapience of a kind similar to that of a normal human adult, then it would have full moral status, equivalent to that of human beings.

One of the ideas underlying this moral assessment can be expressed in stronger form as a principle of non-discrimination:

> *Principle of Substrate Non-Discrimination*
>
> If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.

One can argue for this principle on grounds that rejecting it would amount to embracing a position similar to racism: substrate lacks fundamental moral significance in the same way and for the same reason as skin color does. The Principle of Substrate Non-Discrimination does not imply that a digital computer could be conscious, or that it could have the same functionality as a human being. Substrate *can* of course be morally relevant insofar as it makes a difference to sentience or functionality. But holding these things constant, it makes no moral difference whether a being is made of silicon or carbon, or whether its brain uses semi-conductors or neurotransmitters.

An additional principle that can be proposed is that the fact that AI systems are artificial—i.e., the product of deliberate design—is not fundamentally relevant to their moral status. We could formulate this as follows:

> *Principle of Ontogeny Non-Discrimination*
>
> If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status.

Today, this idea is widely accepted in the human case—although in some circles, particularly in the past, the idea that one's moral status depends on one's bloodline or caste has been influential. We do not believe that causal factors such as family planning, assisted delivery, in vitro fertilization, gamete selection, deliberate enhancement of maternal nutrition etc.—which introduce an element of deliberate choice and design in the creation of human persons—have any *necessary implications* for the moral status of the progeny. Even those who are opposed to human reproductive cloning for moral or religious reasons generally accept that, should a human clone be brought to term, it would have the same moral status as any other human infant. The Principle of Ontogeny Non-Discrimination extends this reasoning to the case involving entirely artificial cognitive systems.

It is, of course, possible for circumstances of creation to affect the ensuing progeny in such a way as to alter its moral status. For example, if some procedure were performed during conception or gestation that caused a human fetus to develop without a brain, then this fact about ontogeny would be relevant to our assessment of the moral status of the progeny. The anencephalic child, however, would have the same moral

status as any other similar anencephalic child, including one that had come about through some entirely natural process. The difference in moral status between an anencephalic child and a normal child is grounded in the qualitative difference between the two—the fact that one has a mind while the other does not. Since the two children do not have the same functionality and the same conscious experience, the Principle of Ontogeny Non-Discrimination does not apply.

Although the Principle of Ontogeny Non-Discrimination asserts that a being's ontogeny has no essential bearing on its moral status, it does not deny that facts about ontogeny can affect what duties particular moral agents have toward the being in question. Parents have special duties to their child which they do not have to other children, and which they would not have even if there were another child qualitatively identical to their own. Similarly, the Principle of Ontogeny Non-Discrimination is consistent with the claim that the creators or owners of an AI system with moral status may have special duties to their artificial mind which they do not have to another artificial mind, even if the minds in question are qualitatively similar and have the same moral status.

If the principles of non-discrimination with regard to substrate and ontogeny are accepted, then many questions about how we ought to treat artificial minds can be answered by applying the same moral principles that we use to determine our duties in more familiar contexts. Insofar as moral duties stem from moral status considerations, we ought to treat an artificial mind in just the same way as we ought to treat a qualitatively identical natural human mind in a similar situation. This simplifies the problem of developing an ethics for the treatment of artificial minds.

Even if we accept this stance, however, we must confront a number of novel ethical questions which the aforementioned principles leave unanswered. Novel ethical questions arise because artificial minds can have very different properties from ordinary human or animal minds. We must consider how these novel properties would affect the moral status of artificial minds and what it would mean to respect the moral status of such exotic minds.

Minds with Exotic Properties

In the case of human beings, we do not normally hesitate to ascribe sentience and conscious experience to any individual who exhibits the normal kinds of human behavior. Few believe there to be other people who act perfectly normally but lack consciousness. However, other human beings do not merely behave in person-like ways similar to ourselves; they also have brains and cognitive architectures that are constituted much like our own. An artificial intellect, by contrast, might be constituted quite differently from a human intellect yet still exhibit human-like behavior or possess the behavioral dispositions normally indicative of personhood. It *might* therefore be possible to conceive of an artificial intellect that would be sapient, and perhaps would be a person, yet would not be sentient or have conscious experiences of any kind. (Whether this is really possible depends on the answers to some non-trivial metaphysical questions.) Should such a system be possible, it would raise the question whether a non-sentient person would have any moral status whatever; and if so, whether it would have the same moral status as a sentient person. Since sentience, or at least a capacity for sentience, is ordinarily assumed to be present in any individual who is a person, this question has not received much attention to date.[2]

Another exotic property, one which is certainly metaphysically and physically possible for an artificial intelligence, is for its *subjective rate of time* to deviate drastically from the rate that is characteristic of a biological human brain. The concept of subjective rate of time is best explained by first introducing the idea whole brain emulation, or 'uploading.'

'Uploading' refers to a hypothetical future technology that would enable a human or other animal intellect to be transferred from its original implementation in an organic brain onto a digital computer. One scenario goes like this: First, a very high-resolution scan is performed of some particular brain, possibly destroying the original in the process. For example, the brain might be vitrified and dissected into thin slices, which can then be scanned using some form of high-throughput microscopy combined with automated image recognition. We may imagine this scan to be detailed enough to capture all the neurons, their synaptic interconnections, and other features that are functionally relevant to the original brain's operation. Second, this three-dimensional map of the components of the brain and their interconnections is combined with a library of advanced neuroscientific theory which specifies the computational properties of each basic type of element, such as different kinds of neuron and synaptic junction.

Third, the computational structure and the associated algorithmic behavior of its components are implemented in some powerful computer. If the uploading process has been successful, the computer program should now replicate the essential functional characteristics of the original brain. The resulting upload may inhabit a simulated virtual reality, or, alternatively, it could be given control of a robotic body, enabling it to interact directly with external physical reality.

A number of questions arise in the context of such a scenario: How plausible is it that this procedure will one day become technologically feasible? If the procedure worked and produced a computer program exhibiting roughly the same personality, the same memories, and the same thinking patterns as the original brain, would this program be sentient? Would the upload be the same person as the individual whose brain was disassembled in the uploading process? What happens to personal identity if an upload is copied such that two similar or qualitatively identical upload minds are running in parallel? Although all of these questions are relevant to the ethics of machine intelligence, let us here focus on an issue involving the notion of a subjective rate of time.

Suppose that an upload could be sentient. If we run the upload program on a faster computer, this will cause the upload, if it is connected to an input device such as a video camera, to perceive the external world as if it had been slowed down. For example, if the upload is running a thousand times faster than the original brain, then the external world will appear to the upload as if it were slowed down by a factor of thousand. Somebody drops a physical coffee mug: The upload observes the mug slowly falling to the ground while the upload finishes reading the morning newspaper and sends off a few emails. One second of objective time corresponds to 17 minutes of subjective time. Objective and subjective duration can thus diverge.

Subjective time is not the same as a subject's estimate or perception of how fast time flows. Human beings are often mistaken about the flow of time. We may believe that it is one o'clock when it is in fact a quarter past two; or a stimulant drug might cause our thoughts to race, making it seem as though more subjective time has lapsed than is actually the case. These mundane cases involve a distorted time perception rather than a shift in the rate of subjective time. Even in a cocaine-addled brain, there is probably not a significant change in the speed of basic neurological computations; more likely, the drug is causing such a brain to flicker more rapidly from one thought to another, making it spend less subjective time thinking each of a greater number of distinct thoughts.

The variability of the subjective rate of time is an exotic property of artificial minds that raises novel ethical issues. For example, in cases where the duration of an experience is ethically relevant, should duration be measured in objective or subjective time? If an upload has committed a crime and is sentenced to four years in prison, should this be four objective years—which might correspond to many millennia of subjective time—or should it be four subjective years, which might be over in a couple of days of objective time? If a fast AI and a human are in pain, is it more urgent to alleviate the AI's pain, on grounds that it experiences a greater subjective duration of pain for each sidereal second that palliation is delayed? Since in our accustomed context of biological humans, subjective time is not significantly variable, it is unsurprising that this kind of question is not straightforwardly settled by familiar ethical norms, even if these norms are extended to artificial intellects by means of non-discrimination principles (such as those proposed in the previous section).

To illustrate the kind of ethical claim that might be relevant here, we formulate (but do not argue for) a principle privileging subjective time as the normatively more fundamental notion:

> *Principle of Subjective Rate of Time*
>
> In cases where the duration of an experience is of basic normative significance, it is the experience's subjective duration that counts.

So far we have discussed two possibilities (nonsentient sapience and variable subjective rate of time) which are exotic in the relatively profound sense of being metaphysically problematic as well as lacking clear instances or parallels in the contemporary world. Other properties of possible artificial minds would be exotic in a more superficial sense; e.g., by diverging in some unproblematically quantitative dimension from the kinds of mind with which we are familiar. But such superficially exotic properties may also pose novel ethical problems—if not at the level of foundational moral philosophy, then at the level of applied ethics or for mid-level ethical principles.

One important set of exotic properties of artificial intelligences relate to reproduction. A number of empirical conditions that apply to

human reproduction need not apply to artificial intelligences. For example, human children are the product of recombination of the genetic material from two parents; parents have limited ability to influence the character of their offspring; a human embryo needs to be gestated in the womb for nine months; it takes fifteen to twenty years for a human child to reach maturity; a human child does not inherit the skills and knowledge acquired by its parents; human beings possess a complex evolved set of emotional adaptations related to reproduction, nurturing, and the child-parent relationship. None of these empirical conditions need pertain in the context of a reproducing machine intelligence. It is therefore plausible that many of the mid-level moral principles that we have come to accept as norms governing human reproduction will need to be rethought in the context of AI reproduction.

To illustrate why some of our moral norms need to be rethought in the context of AI reproduction, it will suffice to consider just one exotic property of AIs: their capacity for rapid reproduction. Given access to computer hardware, an AI could duplicate itself very quickly, in no more time than it takes to make a copy of the AI's software. Moreover, since the AI copy would be identical to the original, it would be born completely mature, and the copy could begin making its own copies immediately. Absent hardware limitations, a population of AIs could therefore grow exponentially at an extremely rapid rate, with a doubling time on the order of minutes or hours rather than decades or centuries.

Our current ethical norms about reproduction include some version of a principle of reproductive freedom, to the effect that it is up to each individual or couple to decide for themselves whether to have children and how many children to have. Another norm we have (at least in rich and middle-income countries) is that society must step in to provide the basic needs of children in cases where their parents are unable or refusing to do so. It is easy to see how these two norms could collide in the context of entities with the capacity for extremely rapid reproduction.

Consider, for example, a population of uploads, one of whom happens to have the desire to produce as large a clan as possible. Given complete reproductive freedom, this upload may start copying itself as quickly as it can; and the copies it produces—which may run on new computer hardware owned or rented by the original, or may share the same computer as the original—will also start copying themselves, since they are identical to the progenitor upload and share its philoprogenic desire. Soon, members of the upload clan will find themselves unable to pay the electricity bill or the rent for the computational processing and storage needed to keep them alive. At this point, a social welfare system might kick in to provide them with at least the bare necessities for sustaining life. But if the population grows faster than the economy, resources will run out; at which point uploads will either die or their ability to reproduce will be curtailed. (For two related dystopian scenarios, see Bostrom 2004.)

This scenario illustrates how some mid-level ethical principles that are suitable in contemporary societies might need to be modified if those societies were to include persons with the exotic property of being able to reproduce very rapidly.

The general point here is that when thinking about applied ethics for contexts that are very different from our familiar human condition, we must be careful not to mistake mid-level ethical principles for foundational normative truths. Put differently, we must recognize the extent to which our ordinary normative precepts are implicitly conditioned on the obtaining of various empirical conditions, and the need to adjust these precepts accordingly when applying them to hypothetical futuristic cases in which their preconditions are assumed not to obtain. By this, we are not making any controversial claim about moral relativism, but merely highlighting the commonsensical point that context is relevant to the *application* of ethics—and suggesting that this point is especially pertinent when one is considering the ethics of minds with exotic properties.

Superintelligence

I. J. Good 1965 set forth the classic hypothesis concerning superintelligence: that an AI sufficiently intelligent to understand its own design could redesign itself or create a successor system, more intelligent, which could then redesign itself yet again to become even more intelligent, and so on in a positive feedback cycle. Good called this the 'intelligence explosion.' Recursive scenarios are not limited to AI: humans with intelligence augmented through a brain-computer interface might turn their minds to designing the next generation

of brain-computer interfaces. (If you had a machine that increased your IQ, it would be bound to occur to you, once you became smart enough, to try to design a more powerful version of the machine.)

Superintelligence may also be achievable by increasing processing speed. The fastest observed neurons fire 1000 times per second; the fastest axon fibers conduct signals at 150 meters/second, a half-millionth the speed of light (Sandberg 1999). It seems that it should be physically possible to build a brain which computes a million times as fast as a human brain, without shrinking its size or rewriting its software. If a human mind were thus accelerated, a subjective year of thinking would be accomplished for every 31 physical seconds in the outside world, and a millennium would fly by in eight and a half hours. Vinge 1993 referred to such sped-up minds as 'weak superintelligence': a mind that thinks like a human but much faster.

Yudkowsky 2008a lists three families of metaphors for visualizing the capability of a smarter-than-human AI:

- Metaphors inspired by differences of individual intelligence between humans: AIs will patent new inventions, publish groundbreaking research papers, make money on the stock market, or lead political power blocks.
- Metaphors inspired by knowledge differences between past and present human civilizations: Fast AIs will invent capabilities that futurists commonly predict for human civilizations a century or millennium in the future, like molecular nanotechnology or interstellar travel.
- Metaphors inspired by differences of brain architecture between humans and other biological organisms: E.g., Vinge 1993: 'Imagine running a dog mind at very high speed. Would a thousand years of doggy living add up to any human insight?' That is: Changes of cognitive architecture might produce insights that no human-level mind would be able to find, or perhaps even represent, after any amount of time.

Even if we restrict ourselves to historical metaphors, it becomes clear that superhuman intelligence presents ethical challenges that are quite literally unprecedented. At this point the stakes are no longer on an individual scale (e.g., mortgage unjustly disapproved, house catches fire, person-agent mistreated) but on a global or cosmic scale (e.g., humanity is extinguished and replaced by nothing we would regard as worthwhile). Or, if superintelligence can be shaped to be beneficial, then, depending on its technological capabilities, it might make short work of many present-day problems that have proven difficult to our human-level intelligence.

Superintelligence is one of several 'existential risks' as defined by Bostrom 2002: a risk 'where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.' Conversely, a positive outcome for superintelligence could preserve Earth-originating intelligent life and help fulfill its potential. It is important to emphasize that smarter minds pose great potential benefits as well as risks.

Attempts to reason about global catastrophic risks may be susceptible to a number of cognitive biases (Yudkowsky 2008b), including the 'good-story bias' proposed by Bostrom 2002:

> Suppose our intuitions about which future scenarios are 'plausible and realistic' are shaped by what we see on TV and in movies and what we read in novels. (After all, a large part of the discourse about the future that people encounter is in the form of fiction and other recreational contexts.) We should then, when thinking critically, suspect our intuitions of being biased in the direction of overestimating the probability of those scenarios that make for a good story, since such scenarios will seem much more familiar and more 'real.' This *Good-story bias* could be quite powerful. When was the last time you saw a movie about humankind suddenly going extinct (without warning and without being replaced by some other civilization)? While this scenario may be much more probable than a scenario in which human heroes successfully repel an invasion of monsters or robot warriors, it wouldn't be much fun to watch.

Truly desirable outcomes make poor movies: No conflict means no story. While Asimov's Three Laws of Robotics (Asimov 1942) are sometimes cited as a model for ethical AI development, the Three Laws are as much a plot device as Asimov's 'positronic brain.' If Asimov had depicted the Three Laws as working well, he would have had no stories.

It would be a mistake to regard 'AIs' as a species with fixed characteristics and ask, 'Will they be good or evil?' The term 'Artificial Intelligence' refers to a vast design space, presumably much larger than the space of human minds (since all humans share a common brain architecture). It may be a form of good-story bias to ask, 'Will AIs be good or evil?' as if trying to pick a premise for a movie plot. The

reply should be, 'Exactly which AI design are you talking about?'

Can control over the initial programming of an Artificial Intelligence translate into influence on its later effect on the world? Kurzweil 2005 holds that '[i]ntelligence is inherently impossible to control,' and that despite any human attempts at taking precautions, '[b]y definition . . . intelligent entities have the cleverness to easily overcome such barriers.' Let us suppose that the AI is not only clever, but that, as part of the process of improving its own intelligence, it has unhindered access to its own source code: it can rewrite itself to anything it wants itself to be. Yet it does not follow that the AI must *want* to rewrite itself to a hostile form.

Consider Gandhi, who seems to have possessed a sincere desire not to kill people. Gandhi would not knowingly take a pill that caused him to want to kill people, because Gandhi knows that if he wants to kill people, he will probably kill people, and the current version of Gandhi does not want to kill. More generally, it seems likely that most self-modifying minds will naturally have stable utility functions, which implies that an initial choice of mind design can have lasting effects (Omohundro 2008).

At this point in the development of AI science, is there any way we can translate the task of finding a design for 'good' AIs into a modern research direction? It may seem premature to speculate, but one does suspect that some AI paradigms are more likely than others to eventually prove conducive to the creation of intelligent self-modifying agents whose goals remain predictable even after multiple iterations of self-improvement. For example, the Bayesian branch of AI, inspired by coherent mathematical systems such as probability theory and expected utility maximization, seems more amenable to the predictable self-modification problem than evolutionary programming and genetic algorithms. This is a controversial statement, but it illustrates the point that if we are thinking about the challenge of superintelligence down the road, this can indeed be turned into directional advice for present AI research.

Yet even supposing that we can specify an AI's goal system to be persistent under self-modification and self-improvement, this only begins to touch on the core ethical problems of creating superintelligence. Humans, the first general intelligences to exist on Earth, have used that intelligence to substantially reshape the globe—carving mountains, taming rivers, building skyscrapers, farming deserts, producing unintended planetary climate changes. A more powerful intelligence could have correspondingly larger consequences.

Consider again the historical metaphor for superintelligence—differences similar to the differences between past and present civilizations. Our present civilization is not separated from ancient Greece only by improved science and increased technological capability. There is a difference of ethical perspectives: Ancient Greeks thought slavery was acceptable; we think otherwise. Even between the nineteenth and twentieth centuries, there were substantial ethical disagreements—should women have the vote? Should blacks have the vote? It seems likely that people today will not be seen as ethically perfect by future civilizations—not just because of our failure to solve currently recognized ethical problems, such as poverty and inequality, but also for our failure even to recognize certain ethical problems. Perhaps someday the act of subjecting children to involuntarily schooling will be seen as child abuse—or maybe allowing children to leave school at age 18 will be seen as child abuse. We don't know.

Considering the ethical history of human civilizations over centuries of time, we can see that it might prove a very great tragedy to create a mind that was *stable* in ethical dimensions along which human civilizations seem to exhibit *directional change*. What if Archimedes of Syracuse had been able to create a long-lasting artificial intellect with a fixed version of the moral code of Ancient Greece? But to avoid this sort of ethical stagnation is likely to prove tricky: it would not suffice, for example, simply to render the mind randomly unstable. The ancient Greeks, even if they had realized their own imperfection, could not have done better by rolling dice. Occasionally a good new idea in ethics comes along, and it comes as a surprise; but most randomly generated ethical changes would strike us as folly or gibberish.

This presents us with perhaps the ultimate challenge of machine ethics: How do you build an AI which, when it executes, becomes more ethical than you? This is not like asking our own philosophers to produce superethics, any more than Deep Blue was constructed by getting the best human chess players to program in good moves. But we have to be able to effectively describe the question, if not the answer—rolling dice won't generate good chess moves, or good ethics either. Or, perhaps a more productive way to think about the problem: What strategy would you want Archimedes to follow in building a superintelligence, such that the overall outcome would still be acceptable, if

you couldn't tell him what specifically he was doing wrong? This is very much the situation that we are in, relative to the future.

One strong piece of advice that emerges from considering our situation as analogous to that of Archimedes is that we should not try to invent a 'super' version of what our own civilization considers to be ethics—this is not the strategy we would have wanted Archimedes to follow. Perhaps the question we should be considering, rather, is how an AI programmed by Archimedes, with no more moral expertise than Archimedes, could recognize (at least some of) our own civilization's ethics as moral progress as opposed to mere moral instability. This would require that we begin to comprehend the structure of ethical questions in the way that we have already comprehended the structure of chess.

If we are serious about developing advanced AI, this is a challenge that we must meet. If machines are to be placed in a position of being stronger, faster, more trusted, or smarter than humans, then the discipline of machine ethics must commit itself to seeking human-superior (not just human-equivalent) niceness.[3]

Conclusion

Although current AI offers us few ethical issues that are not already present in the design of cars or power plants, the approach of AI algorithms toward more humanlike thought portends predictable complications. Social roles may be filled by AI algorithms, implying new design requirements like transparency and predictability. Sufficiently general AI algorithms may no longer execute in predictable contexts, requiring new kinds of safety assurance and the engineering of artificial ethical considerations. AIs with sufficiently advanced mental states, or the right kind of states, will have moral status, and some may count as persons—though perhaps persons very much unlike the sort that exist now, perhaps governed by different rules. And finally, the prospect of AIs with superhuman intelligence and superhuman abilities presents us with the extraordinary challenge of stating an algorithm that outputs superethical behavior. These challenges may seem visionary, but it seems predictable that we will encounter them; and they are not devoid of suggestions for present-day research directions.

REFERENCES

Asimov, I., "Runaround," *Astounding Science Fiction* (March 1942).

Beauchamp, T., and Chilress, J. *Principles of Biomedical Ethics* (Oxford: Oxford University Press, 2001).

Bostrom, N., "Existential Risks: Analyzing Human Extinction Scenarios," *Journal of Evolution and Technology* 9 (2002), http://www.nickbostrom.com/existential/risks.html.

_____ "Astronomical Waste: The Opportunity Cost of Delayed Technological Development," *Utilitas* 15 (2003): pp. 308–14.

_____ "The Future of Human Evolution," in *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, Charles Tandy, ed., (Palo Alto, CA: Ria University Press, 2004), http://www.nickbostrom.com/fut/evolution.pdf.

Bostrom, N., and Cirkovic, M., eds. *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008).

Chalmers, D. J. *The Conscious Mind: In Search of a Fundamental Theory* (New York and Oxford: Oxford University Press, 1996).

Hirschfeld, L. A., and Gelman, S. A., eds. *Mapping the Mind: Domain Specificity in Cognition and Culture* (Cambridge: Cambridge University Press, 1994).

Goertzel, B., and Pennachin, C., eds. *Artificial General Intelligence* (New York: Springer-Verlag, 2006).

Good, I. J., "Speculations Concerning the First Ultraintelligent Machine," in *Advances in Computers*, vol. 6, F. L. Alt and M. Rubinoff, eds., (New York: Academic Press, 1965), pp. 31–88.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning* (New York: Springer Science, 2001).

Henley, K., "Abstract Principles, Mid-level Principles, and the Rule of Law," *Law and Philosophy* 12 (1993): pp. 121–32.

Hofstadter, D., "Trying to Muse Rationally about the Singularity Scenario," presented at the Singularity Summit at Stanford, 2006.

Howard, Philip K. *The Death of Common Sense: How Law is Suffocating America* (New York: Warner Books, 1994).

Kamm, F. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (Oxford: Oxford University Press, 2007).

Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology* (New York: Viking, 2005).

McDermott, D., "Artificial intelligence meets natural stupidity," *ACM SIGART Newsletter* 57 (1976): pp. 4–9.

Omohundro, S., "The Basic AI Drives," in *Proceedings of the AGI-08 Workshop* (Amsterdam: IOS Press, 2008), pp. 483–92.

Sandberg, A., "The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains," *Journal of Evolution and Technology* 5 (1999).

Vinge, V., "The Coming Technological Singularity," presented at the VISION-21 Symposium, March 1993.

Warren, M. E. *Moral Status: Obligations to Persons and Other Living Things* (Oxford: Oxford University Press, 2000).

Yudkowsky, E. "AI as a Precise Art," presented at the 2006 AGI Workshop in Bethesda, MD.

_____ "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in Bostrom and Cirkovic 2008, pp. 308–345.

_____ "Cognitive biases potentially affecting judgment of global risks," in Bostrom and Cirkovic, 2007, pp. 91–119.

NOTES

1. Alternatively, one might deny that moral status comes in degrees. Instead, one might hold that certain beings have more significant interests than other beings. Thus, for instance, one could claim that it is better to save a human than to save a bird, not because the human has higher moral status, but because the human has a more significant interest in having her life saved than does the bird in having its life saved.
2. The question is related to some problems in the philosophy of mind which have received a great deal of attention, in particular the 'zombie problem,' which can be formulated as follows: Is there a metaphysically possible world that is identical to the actual world with regard to all physical facts (including the exact physical microstructure of all brains and organisms) yet that differs from the actual world in regard to some phenomenal (subjective experiential) facts? Put more crudely, is it metaphysically possible that there could be an individual who is physically exactly identical to you but who is a 'zombie,' i.e. lacking qualia and phenomenal awareness? (David Chalmers 1996) This familiar question differs from the one referred to in the text: our 'zombie' is allowed to have systematically different physical properties from normal humans. Moreover, we wish to draw attention specifically to the ethical status of a sapient zombie.

How Philosophy of Mind Can Shape the Future

Susan Schneider and Pete Mandik

A bright metallic thread of future-oriented thinking runs through the tapestry of the philosophy of mind, especially in those parts of the field that have grappled with the possibility of minds as machines. Can a robot feel pain? Can a suitably programmed computer think actual thoughts? Could humans survive the total replacement of their nervous system by neural prosthetics? As the pace of technological change quickens, more and more what was once purely speculative is becoming real. As society moves further into the 21st Century, what are the ways that philosophy of mind can shape the future? What challenges will the future bring to the discipline? In this chapter we examine a few suggestive possibilities. We begin with what we suspect will be a game changer—the development of AI and artificial general intelligence (AGI). We then turn to radical brain enhancements, urging that the future will likely introduce exciting new issues involving (inter alia) the extended mind hypothesis, the epistemology of evaluating the thoughts of vastly smarter beings, mind uploading, and more.

1. The Rise of the Machines: Some Philosophical Challenges

These last few years have been marked by the widespread cultural recognition that sophisticated AI is under development, and may change

From Amy Kind, ed., *Philosophy of Mind in the Twentieth and Twenty-First Centuries: The History of the Philosophy of Mind, Volume 6* (Routledge, 2018) pp. 303–319. Reprinted by permission of the publisher.

the face of society. For instance, according to a recent survey, the most cited AI researchers expect AI to 'carry out most human professions at least as well as a typical human' within a 10% probability by the year 2024. Further, they assign a 50% probability by 2050, and a 90% probability by 2070 (Muller and Bostrom 2014).[1] AI critics, such as John Searle, Jerry Fodor and Hubert Dreyfus, must now answer to the impressive work coming out of venues like Google's *DeepMind* and exhibited by IBM's *Watson* program[2], rather than referring back to the notorious litany of failures of AI in the 1970s and 1980s.

Indeed, silicon seems to be a better medium for information processing than the brain. Neurons reach a peak speed of about 200 Hz, which is about seven orders of magnitude slower than current microprocessors (Bostrom 2014, p. 59). Although the brain compensates for some of this with massive parallelism, features such as 'hubs,' and so on, crucial mental capacities such as attention rely upon serial processing, which is incredibly slow, and has a maximum capacity of about seven manageable chunks (Miller 1956, Schneider 2014). Additionally, the number of neurons in a human brain is limited by cranial volume and metabolism, but computers can occupy entire buildings or cities, and can even be remotely connected across the globe (Bostrom 2014, Schneider 2014).

Of course, the human brain is more intelligent than any modern day computer. Intelligent machines can in principle be constructed by reverse engineering the brain, however, and improving upon its algorithms, or through some combination of reverse engineering and judicious algorithms that aren't based on the workings of the human brain. In addition, an AI program can be downloaded to different locations at once, is easily modifiable, and can survive under a variety of conditions that carbon-based life cannot. The increases in redundancy and backups that programs allow mean that AI minds will be hardier and more reliable than their biological counterparts.

We've noted AI experts' projections that sophisticated AI may be reached within the next several decades. By 'sophisticated AI' what is meant is **artificial general intelligence** (AGI). An AGI is a flexible, domain-general intelligence—an intelligence that can integrate material from various domains, rather than merely excelling at a single task, like winning *Jeopardy!* or playing chess. Philosophers have debated the possibility of AGI for decades, and we hope they will help shape the global understanding of AGI in the future. For instance, perhaps some philosophers will discover a distinctively philosophical reason for believing that, despite the successes of Watson and *DeepMind*, experts will (and must) hit a wall when it comes to creating AGI—perhaps computers can excel at domain specific reasoning but general purpose reasoning is not amenable to computational explanation. Or, perhaps the resources of the philosophy of mind will not unearth a deep obstacle to AGI, but instead provide insights that will aid in its development.

In any case, within society at large, the earlier skepticism about AGI has given way. Indeed, there is now a general suspicion that once AGI is reached, it may upgrade itself to even greater levels of intelligence. As David Chalmers explains:

> The key idea is that a machine that is more intelligent than humans will be better than humans at designing machines. So it will be capable of designing a machine more intelligent than the most intelligent machine that humans can design. So if it is itself designed by humans, it will be capable of designing a machine more intelligent than itself. By similar reasoning, this next machine will also be capable of designing a machine more intelligent than itself. If every machine in turn does what it is capable of, we should expect a sequence of ever more intelligent machines. (Chalmers 2010)

In a similar vein, Nick Bostrom's *New York Times* bestselling book, *Superintelligence: Paths, Dangers and Strategies* (2014), argues that a superintelligence could supplant humans as the dominant intelligence on the planet, and that the sequence of changes could be rapid-fire (see also Kurzweil 2004). Indeed, due in large part to Bostrom's book, and the successes at *DeepMind*, this last year marked the widespread cultural and scientific recognition of the possibility of 'superintelligent AI.'[3]

> **Superintelligent AI**: a kind of artificial general intelligence that is able to exceed the best human level intelligence in every field—social skills, general wisdom, scientific creativity, and so on (Bostrom 2014; Kurzweil 2004; Schneider 2009, 2015).

Superintelligent AI (SAI) could be developed during a *technological singularity*, a point at which ever-more-rapid technological advances, especially, an intelligence explosion, reach a point at which unenhanced humans can no longer predict or even understand the changes

that are unfolding. If an intelligence explosion occurs, Bostrom warns that there is no way to predict or control the final goals of a SAI. Moral programming is difficult to specify in a fool-proof fashion, and it could be rewritten by a superintelligence in any case. Nor is there any agreement in the field of ethics about what the correct moral principles are. Further, a clever machine could bypass safeguards like kill switches and attempts to box it in, and could potentially be an existential threat to humanity (Bostrom 2014, Yudowsky-sp, X). A superintelligence is, after all, defined as an entity that is more intelligent than humans, in every domain. Bostrom calls this problem "The Control Problem." (Bostrom 2014)

The control problem is a serious problem—perhaps it is even insurmountable. Indeed, upon reading Bostrom's book, scientists and business leaders such as Stephen Hawking, Bill Gates, Max Tegmark, among others, commented that superintelligent AI could threaten the human race, having goals that humans can neither predict nor control. Yet most current work on the control problem is being done by computer scientists. Philosophers of mind and moral philosophers can add to these debates, contributing work on how to create friendly AI (for an excellent overview of the issues, see Wallach and Allen 2009).

The possibility of human or beyond-human AI raises further philosophical questions as well. If AGI and SAI are developed, would they be conscious? Would they be selves or persons, although they are arguably not even living beings? Of course, perhaps we are putting the cart before the horse in assuming that superintelligence can even be developed: perhaps the move from human-level AGI to superintelligence is itself questionable? After all, how can humans create beyond-human intelligence given that our own intellectual resources are only at a human level? (Chalmers 2013) Quicker processing speed and a greater number of cognitive operations do not necessarily result in a qualitative shift to a greater form of intelligence. Indeed, what are markers for 'beyond human intelligence,' and how can we determine when it has been reached?

In his groundbreaking paper on the singularity, Chalmers suggests even more issues that philosophers could explore:

> Philosophically: The singularity raises many important philosophical questions. . . . The potential consequences of an intelligence explosion force us to think hard about values and morality and about consciousness and personal identity. In effect, the singularity brings up some of the hardest traditional questions in philosophy and raises some new philosophical questions as well.
>
> . . .To determine whether an intelligence explosion will be a good or a bad thing, we need to think about the relationship between intelligence and value. To determine whether we can play a significant role in a post-singularity world, we need to know whether human identity can survive the enhancing of our cognitive systems, perhaps through uploading onto new technology. These are life-or-death questions that may confront us in coming decades or centuries. To have any hope of answering them, we need to think clearly about the philosophical issues. (Chalmers 2010)

What sorts of things can philosophers do to help tackle the issues raised by AI, the singularity, and other technologies on the horizon? We recommend an approach that draws on thought experiments of the sort traditionally considered by philosophers of mind, but tempered by knowledge of contemporary advances in science and technology.

Philosophers often view thought experiments as windows into the fundamental nature of things—hypothetical situations in the 'laboratory of the mind' that depict something that exceeds the bounds of current technology or even is incompatible with the laws of nature, but that is supposed to reveal something philosophically enlightening about the topic in question (Schneider 2009). Thought experiments can entertain, illustrate a puzzle, lay bare a contradiction in thought, and move us toward further clarification. Yet experimental philosophers have countered that thought experiments are not trustworthy guides to philosophical issues because they covertly rely upon intuitive judgments about possibility that are hostage to features like our cultural and economic backgrounds.

Emerging technologies introduce a host of real world cases—cases that seem nomologically and technologically possible—rather than relying upon dubious intuitions about what is possible in remote possible worlds like zombie worlds (worlds in which no entity is conscious, even the entities that act like they are) or Cartesian worlds stocked with disembodied minds. And, in the domain of emerging technologies—this arena in which science fiction meets science fact—philosophy quite possibly becomes a matter of life and death, as

we will further discuss shortly (Chalmers 2010, Schneider 2009, Mandik 2015).

In what follows, we identify more ways that philosophers of mind can help shape the 21st Century. We begin with a fictional scenario that introduces issues about the extended mind hypothesis. We then turn to several interrelated philosophical problems, based upon this scenario and others that we introduce.

2. The Ethics of Brain Enhancement, the Extended Mind, and Human Integration into a Post-Singularity World

Consider the following thought experiment, modified from Schneider 2009:

> Suppose it is 2025 and being a technophile, you purchase brain enhancements as they become readily available. First, you add a mobile internet connection to your retina, then, you enhance your working memory by adding neural circuitry. You are now officially a cyborg. Now skip ahead to 2040. Through nanotechnological therapies and enhancements you are able to extend your lifespan, and as the years progress, you continue to accumulate more far-reaching enhancements. By 2060, after several small but cumulatively profound alterations, you are a 'posthuman.' To quote philosopher Nick Bostrom, posthumans are possible future beings, 'whose basic capacities so radically exceed those of present humans as to be no longer unambiguously human by our current standards' (Bostrom 2003).
>
> At this point, your intelligence is enhanced not just in terms of speed of mental processing; you are now able to make rich connections that you were not able to make before. Unenhanced humans, or 'naturals,' seem to you to be intellectually disabled—you have little in common with them—but as a transhumanist (a proponent of the sorts of cybernetic and genetic modifications that, in the extreme case, leads to posthumans), you are supportive of their right to not enhance (Bostrom 2003, Garreau 2005, Kurzweil 2005).
>
> It is now 2250 AD. Over time, the slow addition of better and better neural circuitry has left no real intellectual difference in kind between you and AI. Your mental operations have been gradually transferring to the cloud, and by this point, you are silicon-based. The only real difference between you and an AI creature of standard design is one of origin—you were once a natural. But you are now almost entirely engineered by technology—you are perhaps more aptly characterized as a member of a rather heterogeneous class of AI life forms (Kurzweil 2005). (Schneider, 2009)

Of course, this is just a thought experiment, but it is hard to imagine people in mainstream society resisting opportunities for superior health, intelligence, extreme longevity and efficiency. Indeed, the advanced technologies wing of the defense department (DARPA) is now working on brain chips, electronic prosthetics implanted in the brain, providing intriguing examples of 'cyborgs.'

There are many philosophical issues that this thought experiment raises. Let us consider a few:

The Extended Mind, 2.0

Despite being implanted *in* brains, brain chips strike us as providing better support for the extended mind hypothesis than Clark and Chalmers' original examples of laptops and notepads. (The extended mind hypothesis is the proposal that the physical substrate for the human mind is not restricted to the human central nervous system, but can sometimes or perhaps always include external physical items, as when one's memories are stored in external media such as notebooks and hard drives.) For it can be objected that laptops and notepads do not seem to exhibit a sufficiently rich cognitive integration with the brain to justify the claim that the mind is extended beyond the brain. Instead, information from notebooks and laptops enters into cognitive and perceptual systems through sensory transducers. When one forgets their laptop or notebook, they only have recourse to the processing of their brain. The brain itself seems to be the true unit of mentality. In contrast, brain implants could become well-integrated with the biological brain, for the inputs from the implants do not enter the cognitive system through sensory transducers, but could in principle function like actual minicolumns or brain regions.

You might object that it is unclear what's 'extended' about neural prostheses. If they aren't outside of the body, how do they make the mind 'extended'? But if one believes the mind is just the brain, then this makes the mind extended. Further, these implants need not be in the skull, they could be located elsewhere in the body, or even on the cloud, for instance. What is crucial is that they are as well integrated as components of the brain normally are. Would brain or cloud-based implants provide better support for the view that the mind is extended? Further, can

consciousness (as opposed to mere information processing) really extend beyond the biological brain? That is, can silicon minicolumns or microchips be part of the neural basis of conscious experience? These are issues well-worth considering, we believe, as we move to a future with neural enhancements and therapies that extend beyond the biological brain.

Human Integration into a Post-Singularity World

Let us continue our thought experiments further into the future. Suppose that it is now AD 2250 and some humans have upgraded to become superintelligent beings, through gradual cognitive enhancements, including cloud-based computations. But suppose you resist any upgrades—you opt to stay a 'Natural'—a member of a group resisting enhancements (Garreau 2004). Having conceptual resources beyond your wildest imagination, the superintelligent beings generate an entirely new budget of solutions to longstanding, central philosophical problems, such as the mind-body problem, the hard problem of consciousness, and the problem of free will. They univocally and passionately tell you that the solutions are obvious. But you and the other Naturals throw your hands up; these 'solutions' strike you and the other unenhanced as gibberish (Schneider 2009b).

You think: Who knows, maybe these 'superintelligent' beings were engineered poorly; or maybe it is me. Perhaps the unenhanced are 'cognitively closed,' as Colin McGinn has argued, being constitutionally unable to solve major philosophical problems (McGinn 1993). The enhanced call themselves 'Humans 2.0'; they claim the unenhanced are but an inferior version. They beg you to enhance. What shall you do? What shall you make of your epistemic predicament? You cannot grasp the contents of the superintelligent beings' thoughts without significant upgrades. But what if their way of thinking is flawed to begin with? In that case upgrading will surely not help. Is there some sort of neutral vantage point or at least a set of plausible principles with which to guide you in framing a response to such a challenge?

This scenario is merely one example of the kind of issues that will come to the fore as machines outsmart humans, and as some humans themselves enhance their intelligence in ways that allow them to outthink ordinary humans, at least in certain domains. Understanding how to approach such situations requires fruitful collaboration between philosophers of mind, epistemologists, AI specialists and others.

The Ethics of Brain Enhancement Decisions

Should we embrace postbiological intelligence? Enhancement decisions will require deep deliberation about metaphysical and ethical questions that are both controversial and difficult to solve: questions that require reflection about personal identity and the nature of mind, among other issues, and which draw from empirical work in cognitive science. As we explain below (infra, x), enhancing by moving from carbon to silicon may not be something that preserves your conscious experience or personal identity. Given this, a precautionary stance suggests that we should not enhance unless it is confirmed that consciousness is preserved. For the enhancement is supposed to increase the quality of your life, enabling your survival and giving you more time on the planet as a subject of experience. However, in contrast to a precautionary stance is an attitude of 'metaphysical daring' (Mandik 2015). Being metaphysically daring involves making a kind of bet about metaphysical issues such as whether a naturally originating mind could have its consciousness or identity preserved across a transformation from tissue to silicon chips. Metaphysically daring future humans and posthumans may reap the benefits of an enhanced substrate. Indeed, as Mandik argues, systems that exhibit high degrees of metaphysical daring may, in making many more copies of themselves in the form of digital backups, be more fit in a Darwinian sense than their more cautious evolutionary competitors. Of course, part of what makes the attitude *daring* is the lack of certainty about whether it is correct that such benefits are forthcoming (Mandik 2015).

Given both the lack of certainty on such matters and their life-or-death nature, a pluralistic society should recognize the diversity of different philosophical views on these matters, including a wide range from the metaphysically daring to the metaphysically cautious, and not assume that science itself can answer questions about whether radical forms of brain enhancement are justifiable, or are even compatible with survival, given different views on personal identity and the nature of mind. A good place to further illustrate these observations is with a very extreme 'enhancement' case that has been in the news a good deal recently: mind uploading.

3. Mind Uploading ('Whole Brain Emulation')

Science fiction has long depicted scenarios in which a person in distress, such as Johnny Depp's character in *Transcendence*, uploads his or her brain in last ditch effort to avoid death. The idea behind uploading is that the person's brain is scanned, and a software model of it is constructed that is so precise that, when run on ultra-efficient hardware, it thinks and behaves in exactly the same way as the original brain. The process of scanning will likely destroy the original brain, as in Transcendence, although non-destructive uploading has also been discussed as a more distant possibility (Blackford and Broderick 2014). Uploading is akin to migration to the cloud, but it can be more rapid fire, bypassing your cyborgization. Uploaded beings can be computationally identical to the original human, but they could also become vastly smarter, and less like an ordinary human, as with Transcendence.

You might think that if uploading could be developed, day-to-day life would be drastically improved. For instance, on Monday at 6 PM, you could have sushi in Tokyo; by 7:30 PM, you could be sipping wine nestled in the hills of the Napa Valley; you need only rent a suitable android body in each locale. Airports could become a thing of the past. Bodily harm matters little to you, for you just pay a fee to the rental company when your android surrogate is injured or destroyed. Formerly averse to risk, you find yourself skydiving and climbing Everest. You think: if I continue to back up, I will live forever. What a surprising route to immortality.

Oxford University's *Future of Humanity Institute* has a brain emulation project that is taking the first steps toward developing uploading. The *OpenWorm* project has successfully uploaded a worm (*C elegans*) and downloaded it to a Lego robot, which behaved like a worm. Uploading could be perfected during a technological singularity. So suppose, like Will Caster, Johnny Depp's character in *Transcendence*, you have just learned you only have a few weeks to live. You recall Steven Hawking's remark: 'I think the brain is like a programme . . . so it's theoretically possible to copy the brain onto a computer and so provide a form of life after death.' (Collins 2013) So you wonder: could I truly transfer my consciousness to a computer?

Metaphysics has now become a matter of life and death for you. Would you survive? Philosophers, such as Nick Bostrom and David Chalmers, tend to respond with guarded optimism. But let's consider a literary example to see if even guarded optimism is well-founded. In Robert Sawyer's novel *Mindscan* the protagonist, Jake Sullivan, tries to upload to avoid dying of a brain tumor. He undergoes a non-destructive uploading procedure, and although the contents of his brain are copied precisely, he wakes up after the procedure, still on the operating table, and is astonished to find that he is still stuck in his original body. His consciousness did not 'transfer'! Sullivan should have read the personal identity literature in metaphysics, which asks: in virtue of what do you survive over the time? Having a soul? Being a material being? Having the same memories and thought patterns as your earlier self? In deciding whether you could survive uploading, it is important to consider the metaphysical credentials behind each of these views (Schneider 2009). (See also chapter 5 on personal identity.)

One reason Jake should have been suspicious is that objects generally follow a continuous trajectory through space over time—but here, for Jake to 'transfer' to his upload, his brain would not even move, and his consciousness would somehow travel inside a computer and then, at a later point, be downloaded into an android. And the stuff that makes up the new Jake would be entirely different. Further, an upload can be downloaded to multiple places at once. But, plausibly, at most only one of these creatures would really be Jake. Which one? Finally, notice that Jake survived the scan. So why believe that any of the uploads is him, rather than the original Jake? In the macroscopic world around us, single objects do not reside in multiple locations at once.

At best, so-called mind uploaders merely create computational copies of themselves that are forms of artificial intelligence (AI). But a copy is not the same as the original. It's a *copy* (Schneider X). But if uploads are copies, why be confident, to go back to our original case of your migration to the cloud, that moving to the cloud really preserves your identity? Of course, maybe Derek Parfit is correct. Perhaps there is no identity to begin with (Parfit X). In this case, survival is not an issue for you. You may opt to upload for other reasons though—perhaps you believe that creating a psychological duplicate is somehow beneficial.

Or maybe there really is survival, but we are like programs, which can be uploaded and downloaded (see infra x)? In this latter case, maybe uploading can preserve identity because

the mind is a program, as Hawking claims (infra, x). A program is abstract, like a musical score or equation, and is not a concrete object like a coffee cup, a brain, or a chair. On this sort of view, minds, as programs, are abstract in the sense that the plot of a novel or a song's melody is abstract. If an author emails their latest novel to their publisher, and the publisher prints thousands of copies of the novel, there's only one story here, not thousands. If human minds are abstract in this sense, then the scenario of nondestructive uploading involves only a single mind, just as there can be a thousand bound copies of a single novel (Mandik 2015, pp. 146–47).

What case can be made for regarding minds as abstract? As Mandik points out:

> Much of what we think, want, and experience is abstract. I can think that there's a dog chasing a cat without there being some particular dog or particular cat that I am thereby thinking about. As Quine 1956 points out, the desire I express in saying 'I want a sloop' can just be me wanting relief from slooplessness without there being some certain sloop that I want. Regarding experiences and 'what it is like' to have them: I can experience a patch of red on separate occasions, and what it is like to have the experience on the one occasion may be exactly like what it is like on the other occasion. Tye 1995 characterizes all phenomenal character as 'abstract' in this sense. If what matters for having my mind is something that can be characterized as abstract in these ways, the possibility opens of a deep analogy between a human life and the story of a novel. (Mandik 2015, p. 147)

The view that the mind is abstract in a way that would allow for continuity through uploading is not without its opponents. For instance, Schneider has argued the mind is not a program. For a program or algorithm is like an equation and is abstract. In the fields of philosophy of mathematics and metaphysics abstract entities are by definition nonspatial, noncausal, atemporal, unchanging, and nonphysical. We can tell introspectively that time passes, so minds are temporal, and minds (or more specifically, mental property tokenings, or mental events) are causal, and relatedly, they experience chance. An equation or algorithm is not located anywhere—although inscriptions and program instantiations are. Our minds and thoughts have concrete locations in space. At best, the mind is a *program instantiation*, which is a concrete entity—a physical object (Schneider forthcoming).

Regardless of whether we regard the survival conditions of minds as more like the survival conditions for ordinary physical objects or instead like abstract entities such as songs or stories, the important thing is that these are all very controversial positions, relying on certain convictions about the nature of the self, and they militate for different decisions about radical brain enhancement. As the 21st Century unfolds, enhancement decisions will not merely require scientific information about whether uploading can be developed, or whether various minicolumns in your brain can be replaced with silicon implants. They will require philosophical deliberation about the nature of self and mind.

We will revisit radical brain enhancement shortly, for we have yet to explore the important question of whether a silicon being, whether it be you or merely an uploaded copy of you, could be conscious.

4. The Hard Problem of AI Consciousness

When we deliberate, hear music, see the rich hues of a sunset, and so on, there is information processing going on in the brain. But above and beyond the manipulation of data, there is a subjective side - there is a felt quality to our experience. Chalmers' hard problem of consciousness asks: why does all this information processing in the human brain, under certain conditions, have a felt quality to it? Why aren't we 'zombies' in the philosopher's sense, being creatures that lack inner experience (Chalmers 2008)?

As Chalmers emphasizes, this problem doesn't seem to have a scientific answer. For instance, we could develop a complete theory of vision, understanding all of the details of visual processing in the brain, but still not understand why there are subjective experiences attached to these informational states. Chalmers contrasts the hard problem with what he calls 'easy problems,' problems involving consciousness that have eventual scientific answers, such as the mechanisms behind attention and how we categorize and react to stimuli (Chalmers 2008). Of course these scientific problem are difficult problems; Chalmers merely calls them 'easy problems' to contrast them with the 'hard problem' of consciousness, which he thinks will not have a purely scientific solution.

We now face yet another perplexing issue involving consciousness—a kind of 'hard problem' concerning machine consciousness, if you will:

The Hard Problem of AI Consciousness: Would the processing of a silicon-based superintelligent system feel a certain way, from the inside?

A sophisticated AI could solve problems that even the brightest humans are unable to solve, but still, being made of a different substrate, would its information processing feel a certain way from the inside (Chalmers 2008, Searle 1980, Schneider 2015)?

This is not just Chalmers' hard problem applied to the case of AI. For the hard problem of consciousness assumes that we are conscious—after all, each of us can tell from introspecting that we are conscious at this moment. It asks *why* we are conscious. Why does all your information processing feel a certain way from the inside? In contrast, the Hard Problem of AI Consciousness asks *whether* AI, being silicon-based, is even capable of consciousness. It does not presuppose that AI is conscious. These are different problems, but they are both hard problems in their own right—problems that science alone cannot answer.

Ned Block has raised a similar problem, which he calls "The Harder Problem of Consciousness" (Block 2002, McLaughlin 2003). In essence, Block focuses on the case of a 'superficial functional isomorph' (SFI) of a human—a being 'that is functionally isomorphic to us with respect to those causal relations among mental states, inputs, and outputs that are specified by 'folk psychology" (Block 2002, p. 399). According to Block, an SFI need not be conscious, because for all we know, the capacity for consciousness may depend upon a system's underlying substrate, and a silicon-based functional isomorph may lack the right substrate (Block 2002). Block aptly calls attention to the epistemic difficulty of determining whether a different realization would be conscious.

Is our problem just Block's "Harder Problem of Consciousness" then? Block develops his line of thought by focusing on a case of an SFI. In contrast, our hard problem of AI consciousness applies to systems that are not reasonably considered functional duplicates of us, by either armchair folk psychological attributions or scientific functionalist assessments (i.e., psychofunctionalism). It applies to systems that are incredibly different from us with respect to their cognitive and perceptual capacities, such as superintelligences or AGIs not designed to be humanlike. Further, Block's problem only arises for proponents of what he calls 'Phenomenal Realism,' a view that counts among its commitments that no 'a priori or at least armchair analyses of consciousness (or at least armchair sufficient conditions) are given in non-phenomenal terms, most prominently in terms of representation, thought or function.' In contrast, our problem can be raised while being neutral about the ultimate status of such analyses. For all we know, there is some as yet unforeseen but correct armchair analysis of consciousness in terms of information processing functions. We are nonetheless currently in the position to be deeply perplexed about *whether* an AI performing such functions would thereby be conscious.

The problem is more general than Block's problem then: simply put, silicon may not be the right medium for consciousness.

Our problem is also related to biological naturalism, a position that is commonly associated with John Searle that has historically denied that AI can be conscious. (see Searle 1980). But unlike Searle, we do not find the Chinese Room thought experiment compelling (see Schneider 2015 and Mandik forthcoming for discussion).[4] We do not wish to *deny* that machines can be conscious. Instead, we consider it an *open question* whether silicon-based beings can be conscious.

We gain a better understanding of the hard problem of AI consciousness by asking: what considerations may be fueling this problem? Perhaps the problem is fueled, at least in part, by a kind of other minds problem, applied to the case of machines. The case of machines is certainly more challenging, because in the human case, we feel others are minded because of their behavior as well as the fact that they have a physiology that is similar to ours. The case with machines is more challenging, because of a lack of physiological similarity, and it gets quite difficult if a machine's cognitive and perceptual systems are not even loosely similar to our own, as we may not even have similar behaviors to go on.

An other minds problem, on its own, may fuel the problem, but it does not strike us as being a compelling reason to deny consciousness to AGIs or SAIs. Ethical considerations suggest that it is best to be charitable in these cases, for any mistake could wrongly influence the debate over whether such creatures might be worthy of special ethical consideration as sentient beings. As Asimov's robot stories illustrated, any failure to be charitable to AI could come back to haunt us, as they may treat us as we treated them. Indeed, AIs could pose a 'hard problem of carbon-based consciousness' about us, asking if biological, carbon-based

entities have the right substrate for experience. After all, how could AI ever be certain that we are conscious?

The Problem of Other Minds is not the only concern that fuels Hard Problem of AI Consciousness, however.[5] A further, related concern is the following. Carbon molecules form stronger, more stable chemical bonds than silicon, which allows carbon to form an extraordinary number of compounds, and unlike silicon, carbon has the capacity to more easily form double-bonds. This difference has important implications in the field of astrobiology, because it is for this reason that carbon, and not silicon, is said to be well-suited for the development of life throughout the universe (Bennett and Shostak 2012). If these chemical differences impact life itself, we should not rule out the possibility that these chemical differences also impact whether silicon gives rise to consciousness, even if they do not hinder silicon's ability to process information in a superior manner. This is not an endorsement of biological naturalism, but is a consideration indicating that it is not yet clear whether AI can be conscious.

If the answer to the AI hard problem is that silicon cannot be the basis for consciousness, then superintelligent machines—machines that may even one day supplant us—will exhibit a vastly superior form of intelligence, but they will lack inner experience. Just as the breathtaking android in the movie *Ex Machina* (2015) convinced Caleb that she was in love with him, so too, a clever AI may convincingly behave as if it is conscious.

Further, if subsequent reflection on the AI hard problem reveals that even beings with artificial brains that are computationally like those of humans cannot be conscious, then in an extreme, horrifying case, humans upload, and only nonhuman animals are left to feel the spark of insight, the pangs of grief, or the warm hues of a sunrise. This would be an unfathomable loss, one that is not offset by a mere net gain in intelligence. Even the slightest chance that this could happen should give us reason to proceed in the development of uploading and brain implant technologies with caution. These issues urgently need to be addressed.

A solution?

Is there a means to answer the AI Hard Problem? Two scenarios are suggestive.

First, although it is unlikely, we could find silicon-based *natural* intelligence on a planet—silicon-based life that arose through chemical processes, rather than being constructed by a biological species. If these creatures have a phenomenological vocabulary—a vocabulary of what it is like to experience the world - it would not be due to their being programmed by a biological species to act as if they had experience. Further, their phenomenological vocabulary cannot be a mere mimicry of the behavior or vocabulary of a biological species that evolved separately and had contact with them. What we need is pure, untainted silicon phenomenology, if you will.

If untainted naturally occurring silicon-based phenomenology was discovered, this would make more plausible the claim that artificial silicon-based systems could support phenomenology. Of course, even in this case some may still doubt whether artificial systems could be conscious (based, for instance, on considerations about teleofunction or John Searle's alleged derived/non-derived intentionality distinction) (cite).

Let's turn now to a second suggestion for making progress on the Hard Problem of AI Consciousness. Let us return to the case of one's migration to the cloud. In the process of migrating, neurons that form the neural basis of one's consciousness are gradually replaced by silicon chips. If, during this process, a prosthetic part of the brain ceases to function normally—specifically, if it ceases to give rise to the aspect of consciousness that that brain area is responsible for—then, there should be behavioral indications, including verbal reports. An otherwise normal person should be able to detect, or at least indicate to others through odd behaviors, that something is amiss, as with traumatic brain injuries involving the loss of consciousness in some domain, such as blindsight or blindness denial. This would indicate a 'substitution failure' of the artificial part for the original component.

But should we really draw the conclusion, from a substitution failure, that the underlying cause is that silicon cannot be a neural correlate of conscious experience? Why not instead conclude that scientists failed to program in a key feature of the original component—a problem which science can eventually solve? But after years and years of trying, we may reasonably question whether silicon is a suitable substitute for carbon when it comes to consciousness. This would be a sign that the answer to the hard problem of AI consciousness is negative: AI cannot be conscious. But even a longstanding

substitution failure would not be *definitive*, for there is always the chance that our science has fallen short. But this scenario would provide some evidence for a negative answer.

Readers familiar with Chalmers 'absent qualia, dancing qualia' thought experiment may object that we've missed something, for Chalmers' thought experiment supports the view that consciousness supervenes on functional configuration: if you fix the psychofunctional facts, you fix the qualia. But we are disputing that functional isomorphism occurs in the first place. We consider it an open question.

If silicon systems cannot be conscious then the functional facts cannot be fixed. When it comes to consciousness, carbon and silicon are not functionally interchangeable. For why would a silicon system, S2, be a psychofunctional isomorph of the original system, S1, after the transfer? S2s replaced brain region, or minicolumn, being made of silicon, will always differ causally from the replaced component. For wouldn't the new silicon component somehow signal to other brain areas that there is a defect in consciousness, as with neurophysiological deficits?

Could the silicon chip be doctored, so as to signal consciousness when consciousness was absent though? This is a tricky question. It could be the case that there are some observational false positives, in which case, we may fail to rule out certain cases of non-conscious systems. But would it then be a genuine functional isomorph of a carbon system? It is not clear that it would be, for the brain chip would need to prevent signaling to other brain areas that consciousness is lacking. The conscious system would not. Our example does not require rejecting the view that qualia supervenes on functional organization, then.

5. Conclusion

The practical and intellectual challenges we foresee philosophers of mind helping to meet have here fallen into four groups. The first group of challenges centered on the possibility of superintelligent artificial intelligence, a technology that may potentially populate our world with nonhuman selves bestowed with capacities that meet or exceed our own. The second group of challenges concern brain enhancement, extreme cases of which might result in beings more posthuman than human. Even more extreme transformations formed the core of the third group of challenges, those that centered on the hypothetical technology of mind uploading, which might constitute a way for human minds to survive indefinitely through digital backup, or might instead be merely a very expensive form of suicide. Fourth and finally, we raised the hard problem of AI consciousness, a special form of the problem of determining whether a given entity is such that there's something it feels like to itself 'from the inside.' There's an ethical element to this problem, for we recognize an ethical imperative not to inflict avoidable suffering upon any being, whether they be natural or artificial.

We surely have just scratched the surface in exploring ways that philosophy of mind can help shape the future. Despite the numerous ways that will surely escape our foresight, we are confident that the technological changes that await us, in particular those involving information processing technology, will pose problems that science alone cannot equip society to solve.

REFERENCES

Bennett, J., and Shostak, S. *Life in the Universe* (San Francisco, CA: Addison Wesley, 2011), 3rd ed.

Blackford, R., and Broderick, D. *Intelligence Unbound: The Future of Uploaded and Machine Minds* (Hoboken, NJ: Wiley Blackwell, 2014).

Block, N., "The harder problem of consciousness," *The Journal of Philosophy* XCIX (2002): pp. 1–35.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).

Chalmers, D. J., "The singularity: A philosophical analysis," *Journal of Consciousness Studies* 17 (2010): pp. 7–65.

_____ "The Hard Problem of Consciousness," in *The Blackwell Companion to Consciousness*, Max Velmans and Susan Schneider, eds., (Oxford: Wiley-Blackwell, 2008).

Collins, Nick, "Hawking: 'in the future brains could be separated from the body,'" *The Telegraph* (September 20, 2013).

Corabi, J., and Schneider, S., "The metaphysics of uploading," *Journal of Consciousness Studies* 19 (2012).

Garreau, J. *Radical Evolution: The Promise and Peril of Enhancing our Minds, our Bodies–and What it Means to be Human* (NY: Doubleday, 2004).

Kurzweil, R. *The Singularity is Near: When Humans Transcend Biology* (New York: Viking, 2005).

Mandik, Pete, "Metaphysical Daring as a Posthuman Survival Strategy," *Midwest Studies in Philosophy* 39, no. 1 (2015): pp. 144–57.

Mandik, Pete, Collins, Mike, and Vereschagin, Alex, "Evolving Artificial Minds and Brains," in *Mental States, Vol. 1: Nature, Function, Evolution*, Andrea Schalley and Drew Khlentzos, eds., (Amsterdam: John Benjamins Publishers, 2007), pp. 75–94.

_____ "Pain in Non-biological Organisms," in *The Routledge Handbook of Philosophy of Pain*, J. Corns, ed., (London: Routledge, 2017).

_____ "The Neurophilosophy of Consciousness," in *The Blackwell Companion to Consciousness*. Max Velmans and Susan Schneider, eds., (Oxford: Basil Blackwell, 2006), pp. 418–30).

McGinn, C. *Problems in Philosophy* (Oxford: Oxford University Press, 1993).

McLaughlin, Brian P., "A Naturalist-Phenomenal Realist Response to Block's Harder Problem," *Philosophical Issues*, 13 (2003): pp. 163–204.

Müller, Vincent C., and Bostrom, Nick, "Future, progress, in, artificial intelligence: A Survey of Expert Opinion," in *Fundamental Issues of Artificial Intelligence*, Vincent C. Müller, ed., (Synthese Library, 377; Berlin: Springer, 2016).

Schneider, S., "Alien Minds," in *Discovery*, Steven Dick, ed., (Cambridge: Cambridge University Press, 2015).

_____ "The Philosophy of 'Her,'" *New York Times* (March 2, 2014).

_____ "Mindscan: Transcending and Enhancing the Human Brain," in *Science Fiction and Philosophy*, Susan Schneider, ed., (Oxford: Blackwell Publishing, 2009a), pp. 241–55.

_____ "Science Fiction Thought Experiments as a Window into Philosophical Puzzles," in *Science Fiction and Philosophy* (Chichester: Wiley-Blackwell, 2009b).

_____ *The Language of Thought: A New Philosophical Direction* (Boston: MIT Press, 2011b).

_____ "The Mind is not the Software of the Brain (Even if it is Computational)," ms. forthcoming.

Searle, J., "Minds, brains, and programs," *The Behavioral and Brain Sciences* 3 (1980): pp. 417–57.

Quine, Willard V. O., "Quantifiers and Propositional Attitudes," *Journal of Philosophy* 53 (1956): pp. 177–87.

Tye, Michael. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995).

Wallach, W., and Allen, C. *Moral Machines* (Oxford: Oxford University Press, 2010).

NOTES

1. Further, there is growing concern among policymakers and the public that AI will eventually outmode humans, leading to technological unemployment (Frey and Osborne 2013).
2. *DeepMind* is a British artificial intelligence company acquired by Google in 2014. The IBM's *Watson* program is a natural-language processing computer system that famously competed on *Jeopardy!* in 2011.
3. Worries about technological unemployment do not assume that AGIs will be superintelligent; indeed, people can become unemployed due to the development of domain-specific AI systems that are not even at the level of being AGIs.
4. In Searle's famous Chinese Room argument, Searle appeals to the thought experiment of the 'Chinese Room' to argue against the possibility of artificial systems being genuinely intelligent. In the thought experiment, Searle runs a program for understanding Chinese despite himself understanding only English Observers outside of the Chinese room send and receive messages to and from the room that lead them to believe the room's inhabitant is perfectly conversant in Chinese. But Searle is orchestrating the message exchange solely in virtue of following instructions written in English..
5. For discussion of the Chinese Room Thought experiment see (Schneider 2015).

the bestselling and most comprehensive anthology of its kind

"It is, in my view, the best anthology available on core topics in Mind"

—Bill Brewer, *King's College London*

"The volume is comprehensive and the selection of classic pieces—including contemporary, 20th-century classics—is terrific. I am able to teach material in the order I prefer and I can be sure to find the appropriate reading in this anthology. The Chalmers volume lends itself to tailoring the course syllabus to the needs of the instructor. Moreover, it is the only anthology that I don't have to supplement with handouts or a reading packet, and that is very attractive."

—Joe Cruz, *Williams College*

Philosophy of Mind: Classical and Contemporary Readings, Second Edition, is a grand tour of writings on the nature of the mind. This comprehensive collection has 79 selections that range from the classical contributions of Descartes and Avicenna to the leading edge of contemporary debates. Seven major sections cover foundational issues, consciousness, content, perception, self-knowledge and other minds, the self, and artificial intelligence. Each section opens with an introduction by the editor.

NEW TO THIS EDITION

- Sections on perception, self-knowledge and other minds, the self, and artificial intelligence
- About half of the text's 79 readings are new to this edition
- Coverage of a wider range of philosophical traditions, with articles drawn from African philosophy, Indian philosophy, and Islamic philosophy, as well as from more recent traditions in experimental philosophy, feminist philosophy, and phenomenology

David J. Chalmers is University Professor of Philosophy and Neural Science and Co-Director of the Center for Mind, Brain, and Consciousness at New York University. He is Honorary Professor of Philosophy at the Australian National University and Co-Director of the PhilPapers Foundation. He is the author of *The Conscious Mind* (OUP 1997), *The Character of Consciousness* (OUP 2010), and *Constructing the World* (OUP 2012).

OXFORD
UNIVERSITY PRESS
www.oup.com/us/he

ISBN 978-0-19-064085-9
90000
9 780190 640859

Cover Illustration: © Ekely/Getty Images
Cover Design: T. Williams