TOWARD A THEORY OF CONSCIOUSNESS

by

David J. Chalmers

Center for Research on Concepts and Cognition
Indiana University
Bloomington, IN 47408.
dave@cogsci.indiana.edu

(As of September 1993)
Department of Philosophy
Washington University
St. Louis, MO 63130.

November 29, 1993

Contents

Acknowledgments

My first thanks go to Doug Hofstadter. It was his writing that introduced me to the mysteries of the mind when I was young. As my advisor the last few years, he has been tolerant and supportive as my interests have veered all over the map. Although he disagrees with much of what I say herein and finds it too philosophical to boot, he has remained interested and we have had a number of absorbing conversations about the subject matter. He has also provided a comfortable and stimulating research environment in the Center for Research on Concepts and Cognition, without whose resources this work might have taken much longer.

Many thanks also to my committee co-chair, Mike Dunn, who has been helpful to me in all sorts of ways. Rob Goldstone and Tim van Gelder, the other committee members, have made numerous helpful remarks.

I first became excited by consciousness and the mind–body problem as an undergraduate studying mathematics at the University of Adelaide. Conversations with a number of people, especially Paul Barter, Jon Baxter, Ben Hambly, and Paul McCann, helped form my ideas. Even at the time the subject seemed to be about as fascinating a problem as there could be. It seemed faintly unreasonable that somebody could be occupied full-time thinking about something that was so much fun.

Later, as a graduate student at Oxford, I found that the mind was always occupying my thoughts where mathematics should have been, and I decided to switch fields. My advisor, Michael Atiyah, was thoroughly supportive in this difficult time, as were Michael Dummett in philosophy and Robin Fletcher in Rhodes House. I also had some interesting conversations about consciousness with Colin McGinn and Kathy Wilkes.

After some correspondence with Doug Hofstadter and a visit to Indiana, I decided that Indiana University, with the double attraction of its new cognitive science program and its fine philosophy department, was the best place for me. I have had a good four years here. There has been enough going on that I have often been distracted; it has been easy to get seduced into working on such topics as connectionist networks, genetic algorithms, and the foundations of artificial intelligence. Consciousness has remained my first and greatest love, however, and the last year of devoting myself to the subject full-time has been perhaps the most exciting intellectual period of my life.

For most of my time here, my work on consciousness has been a solitary project, but recently interest in consciousness has mushroomed, and I have had many interesting discussions about the topic. In particular I would like to thank Gregg Rosenberg and Jerry Seligman, who are as interested in the subject as I am and whose views have had much effect on my own. The consciousness discussion group in the back room at Nick's also provided many enjoyable Monday afternoon conversations; thanks especially to Jim Hettmer and Audrey Bancroft for getting this group going. Fellow members of the Center for Research on Concepts and Cognition, including Bob French, Liane Gabora, Jim Marshall, Gary McGraw, Melanie Mitchell, and David Moser, have all had interesting things to say, as have visitors such as Morten Christiansen, Eric Dietrich, and Peter Suber. I would also like to thank many others in the philosophy department and elsewhere for stimulating discussion.

I should also like to acknowledge much valuable correspondence on the topic, most of it electronic. Among the many correspondents to whom I am indebted are Ned Block, Dan Dennett, Edward Paul Faith, Guven Guzeldere, Terry Horgan, Drew McDermott, Brian McLaughlin, Calvin Ostrum, and Aaron Sloman.

My many intellectual debts should be obvious from the text. I developed my initial views on consciousness largely on my own, but since then I have developed my views and read the literature in parallel. The two influences combine by a subtle feedback effect, so it is often difficult to credit a particular argument herein to one source or the other. Thomas Nagel has done more than anyone to emphasize the perplexing nature of the problem of consciousness; many of the points I make in Chapter 3 have been anticipated by him, although I do not share his pessimism. Dan Dennett has provided a worthy stalkinghorse throughout. The metaphysical framework that I develop in Chapter 2 owes much to the work of Terry Horgan and David Lewis, among others. The work of Saul Kripke provided the impetus to develop a framework for dealing with pervasive problems about meaning and necessity. Sydney Shoemaker's work influenced the discussion on phenomenal judgments in Chapters 5 and 6 in a deeper way than is obvious from the text. Of those who have worked on consciousness from the standpoint of cognitive science, Ray Jackendoff's approach is perhaps closest to my own; his book on consciousness is the one I most admire. Doug Hofstadter's way of thinking about the mind has infused my work, even when I have come to conclusions vastly different from his. Ned Block's, Frank Jackson's, and John Searle's influential discussions of the problems of consciousness have all played a significant role. Finally, I would also like to acknowledge some less well-known but insightful work by Robert Kirk, Joe Levine, and Bill Seager, which anticipates the work herein at a number of points.

Thanks to Helga Keller for all her administrative help. Thanks to Lisa Thomas for a book on zombies and for moral support. Thanks to all three of my parents for their parenting, and for their encouragement of my intellectual ambitions. Thanks to the qualia produced by my bookcases and my socks for keeping me going. And thanks to everybody else.

Preface

Consciousness is the biggest mystery. It is probably the largest outstanding obstacle in our quest for a scientific understanding of the universe. The science of physics is not yet complete, but it is well-understood. The science of biology has explained away many of the mysteries surrounding the nature of life. There are many gaps in our understanding of these fields, but they do not seem intractable. We have some idea of what a solution that would fill these gaps might look like; it is just a matter of coming up with a theory that gets the details right.

In the science of the mind, things are not so rosy. To be sure, much progress has been made in the explanation of human behavior, and of the cognitive processes that lead to that behavior. It seems clear enough in principle that behavior can be explained in physical or functional terms, and that there is no need to postulate a ghost in the machine to do any causal work. We do not have many good detailed theories of cognition, to be sure, but there are few problems of *principle*; the details cannot be far off. Consciousness, however, is as much of a problem as it ever was. It still seems utterly mysterious that the causation of behavior should be accompanied by conscious experience. We do not just lack a detailed theory; we have little idea what a theory of consciousness would even look like.

It seems almost certain that consciousness arises from physical systems such as brains, but we have little idea how it so arises, or why it exists at all. How could a physical system like a brain also be an *experiencer*? Why should there be *something it is like* to be such a system? Currently, we have little idea how to answer these questions. Present-day scientific theories hardly touch the really difficult questions about consciousness. In the pervasive inter-level explanatory structure that connects physics, chemistry, biology, psychology, and higher-level phenomenon, consciousness sticks out like a sore thumb by its absence.

All this means that the study of consciousness is difficult, but it also makes it exciting. In other domains, the overall shape of our worldview is becoming fixed. To be sure, there are likely to be minor revolutions in our understanding of physics, biology, and psychology, but in these fields, we have probably at least got the basics right. With consciousness, we do not even have the basics down; we are entirely in the dark about how consciousness fits into the natural order. This means that a correct theory of consciousness is likely to affect our conception of the universe more profoundly than any other new scientific development. Consciousness is both fundamental and unexplained; this makes for a potent cocktail.

In this work, I do not pretend to solve the problem of consciousness, but I try to rein it in. I outline a general explanatory and ontological framework within which the many problems can be addressed, and I argue in detail that a certain kind of reductive explanation cannot work. I develop a picture of what a theory of consciousness must look like, if there is to be such a theory at all. In the second half of this work, I go some way toward fleshing out the details of such a theory, by putting strong constraints on its form. The only full-fledged "theory" of consciousness that I put forward is quite speculative and tentative, but at least it gives a picture of what such a theory might look like. There is much still to do, but I hope to have pointed in the right direction.

In developing my account of consciousness, I have subjected myself to a number of constraints. The first is to *take consciousness seriously.* The easiest way to develop a "theory" of consciousness is to deny its existence, or to redefine the phenomenon in need of explanation as something it is not. Some say that consciousness is an "illusion", but I have little idea what this could even mean. It seems to me that we are surer of the existence of conscious experience than we are of anything else in the world. I have tried hard at times to convince myself that there is really nothing there, that conscious experience is empty, an illusion. There is something seductive about this notion, which philosophers from Wittgenstein to Dennett have exploited, but in the end it is utterly unsatisfying. I find myself absorbed in an orange sensation, and *something is going on.* There is something that needs explaining, over and above my dispositions to react; there is the *experience.*

While I suppose the position that conscious experience does not exist is a coherent one, it is so implausible that it would require an argument more sophisticated than any I have ever seen. In the absence of such an argument, I have taken it as given that conscious experience exists.

Secondly, I have constrained myself to take current scientific theories at their word in the domains where they have authority, but I have not been afraid to go out on a limb in areas where scientists' opinions are as ungrounded as anyone else's. For instance, current physics seems to indicate strongly that the physical domain is causally closed, and I have no authority to dispute this; but if a physicist hopes that consciousness should be explainable in physical terms, this is merely a hope ungrounded in current theory, and the question remains open. The same goes for cognitive science: current theory provides excellent grounds for believing that behavior should be explainable in physical or functional terms, but it says very little about consciousness; cognitive scientists' speculation about consciousness deserve no special authority. In a similar way, quantum theory provides an formidable theory for predicting the results of certain measurements, which I would not dream of doubting; but the interpretation of what is going on in the real world behind these measurement is as mysterious to physicists as to the rest of us, so I have not been afraid to develop my own speculation here.

The third constraint is that consciousness is viewed as a natural phenomenon, and must ultimately fall under the sway of natural laws. It follows that there is *some* correct scientific theory of consciousness, whether or not we can arrive at such a theory. That consciousness is a natural phenomenon seems indisputable; it is an extraordinarily salient part of nature, and cries out for explanation. Its many differences with other natural phenomenon should not be led to obscure this fundamental fact. We have every reason to believe that natural phenomenon fall under the sway of fundamental scientific laws; it would be strange if consciousness did not, and it would make the development of a theory of consciousness almost impossible. I have therefore assumed, almost as a methodological assumption, that it must. This is not to say that the scientific laws concerning consciousness will be just like laws in other domains; we will see that they may be significantly different in kind.

The problem of consciousness lies uneasily at the border of science and philosophy. I would say that it is properly a scientific subject matter, being a contingent natural phenomenon like motion, life, and cognition, calling out for explanation in the way that these do. But it is not open to investigation by the usual scientific methods. Familiar scientific methodology has problems even getting a grip on the problem of consciousness, not least because of all the problems we have in *observing* the phenomenon intersubjectively. Outside our own case, data are hard to come by. This is not to say that no external data can be relevant, but we first have to arrive at a coherent philosophical understanding before we can even justify the relevance of that data. In the meantime, the problem of consciousness is a scientific problem that requires philosophical methods of understanding before we can even get off the ground.

In this work I come to conclusions that some may see as antithetical to a respectable contemporary scientific worldview. For instance, I repudiate the possibility of a reductive explanation of consciousness, of the kind that has been successful for other natural phenomena; and I even advocate dualism of a certain variety. It is important to understand two things; first, these are *conclusions* in the strongest sense, and not premises. I have little in the way of spiritual or religious inclination, and have generally thought of myself as quite hard-headed in my approach to these matters. Temperamentally, I am strongly inclined toward materialist reductive explanation. For a number of years I hoped for such a theory; when I eventually gave up the "materialist" label, it was quite reluctantly. It eventually seemed plain to me that these conclusions are simply forced on anyone who wants to take consciousness seriously. Even on a hard-headed scientific worldview, if we want a theory of the phenomenon at all, we are forced to go beyond the resources that reductive explanation provides us with.

Second, giving up on reductive explanation does not mean giving up the quest for a naturalistic explanation of consciousness. While non-materialist views have often been associated with the notion that consciousness is transcendental and beyond our understanding, I do not think that any such conclusion follows. The theories of consciousness that I advocate are all quite law-governed in their shape, and quite compatible with current scientific theories; it is just they they expand the ontology involved. To those who suspect that science requires materialism, I ask that you wait and see. While my initial embracing of these conclusions was quite reluctant, I have grown almost happy with them. Our scientific worldview need not be overthrown; it only needs to be expanded. Even this may be all for the best; it makes the universe a more interesting place.

This work consists largely in two parts, a negative part and a positive part, with some preliminary material preceding the main arguments, and with some applications presented in an appendix. The preliminaries begin in Chapter 1, which is an introduction to the problem of consciousness, and an account of the subtle relationship between notions of consciousness and other mental motions. As has often been noted, "consciousness" is a highly ambiguous term. In this chapter, I try to find some systematicity in that ambiguity, and to draw out the relevant sense of the term, the sense in which consciousness is really *interesting*.

Chapter 2 develops the metaphysical and explanatory framework in terms of which most of the later discussion is framed. This centers on the notion of supervenience, and develops the relevance of various kinds of supervenience for explanation and ontology. I give an account of the sort of reductive explanation that is pervasive in science, spell out in relation to supervenience, and give arguments as to why it should be so pervasive. A phenomenon is reductively explainable precisely when it is *logically supervenient* on the physical facts, and there are good reasons to believe that *almost* every phenomenon is logically supervenient in this way.

Chapter 3 constitutes the focus of the negative part. In this chapter I argue that consciousness almost uniquely is not supervenient in this sense, and so is not reductively explainable. I illustrate this conclusion with an account of various purported reductive explanations of consciousness that have been put forward by cognitive scientists, neuroscientists, and others. Chapter 4 takes things further by arguing that this failure of logical supervenience has significant ontological consequences: it implies the falsity of materialism and the truth of a variety of dualism, according to which consciousness is merely *nomically* supervenient on the physical.

I suspect that the negative material will provoke the most reaction in readers, but I do not view it as the primary contribution of this work. Initially, I saw the negative material as a chore to be got out of the way before the interesting stuff could be said. Eventually it took on a life of its own as I realized the need to develop the argument as thoroughly as possible, and it even turned out to be extremely enjoyable to write, but it is still essentially preliminary. My real goal is to develop a theory of consciousness, not to knock down the theories of others.

Ideally, much of the positive material will be of value even to those who do not accept my negative conclusions. I developed much of this before I came to accept the negative views put forward earlier—particularly the dualism of Chapter 4—so it should stand independently to a large extent. The material on structural coherence in Chapter 6, on the relationship between consciousness and functional organization in Chapter 7, and on the deep connection between consciousness and information in the appendix puts forward arguments and conclusions that even many materialists might be sympathetic with. Although my discussion in these chapters is often influenced by earlier chapters, the material should be straightforwardly adaptable to a different framework.

The positive part begins with the end of Chapter 4, which outlines the shape that a non-reductive theory of consciousness might take. The next few chapters put some flesh on these bones. Chapter 5 and 6 discuss the complex and interesting relationship between consciousness and our *judgments* about consciousness. We will see in Chapter 5 that these pose perhaps the greatest problem for a nonreductive theory of consciousness: almost paradoxically, our judgments about consciousness are reductively explainable even though consciousness itself is not. The paradox can be defused, however, and considerations about the subtle relationship between consciousness and our judgments about it can provide strong scaffolding for a theory of consciousness to be built on.

In Chapter 6 I consider this relationship in more detail, focusing on the systematic *coherence* between conscious experience and cognitive states. I investigate various kinds of coherence, culminating in an extremely useful principle of *structural coherence* between consciousness and its functional correlates. I also try to resolve some epistemological difficulties that the non-reductive view might be thought to pose.

Chapter 7 discusses the relationship between consciousness and functional organization, using thought-experiments to argue that fine-grained patterns of causal organization fully determine conscious experience in an individual. According to the "Absent Qualia" and "Inverted Qualia" hypothesis, our functional organization might be instantiated without experience at all, or with a different sort of experience. I develop "Fading Qualia" and "Dancing Qualia" thought-experiments to provide a *reductio* of these hypotheses; they are logically possible, but almost certainly empirically impossible. It follows our theories of consciousness are constrained to have conscious experience depend on functional organization, which is a significant advance on unconstrained physical dependence.

The appendix summarizes some future work that will extend the work I have put forward here. A first line of future work will conclude the positive development of a theory of consciousness by putting forward a detailed speculative theory that meets the constraints outlined so far. Two other lines of future work will apply the conclusions I have reached about consciousness to central issues in artificial intelligence and quantum mechanics, arguing for what Searle has called "Strong AI"—the thesis that there exists a class of algorithms such that implementation of an algorithm in that class suffices for consciousness—and for a version of Everett's "no collapse" view of quantum theory.

I have tried my best to make this work accessible to non-philosophers as well as to philosophers. I have used technical terminology only when necessary, and those instances I have introduced and motivated it as well as I can. The most important technicalities are in the material on supervenience in section 2.1; once these are clear, everything else should follow without too much difficulty. To philosophers, some passages may seem laborious or repetitive, but I have valued clarity over pyrotechnics. Some parts of the discussion, such as those in which I engage the existing literature, have necessarily been philosophically technical. I have generally marked those sections; those without a strong philosophical background can skip these sections without losing the thread.

For those who want the short version of this work, most of the chapters should make sense in isolation, although each presupposes those preceding to a limited extent. For the dime tour, glance at Chapter 1, read the material on supervenience in 2.1 as necessary background material, and glance at the rest of Chapter 2. Then read all of Chapter 3 for the central arguments against reductive explanation, and sections 4.1 and 4.7 for the central considerations about dualism. Of the positive material, Chapter 7 is perhaps the most self-contained as well as the most fun; some readers without an inclination for the negative material might like to read this chapter first to whet their appetites.

A couple of notes for philosophers. First, the literature on consciousness is largely a mess, with seemingly independent strands talking about related issues without making contact with each other. I have attempted to impose some structure on the sprawl by providing a unifying framework in which the various metaphysical and explanatory issues become clear. Much of the discussion in the literature can be translated into this framework without loss. Where it cannot be, this is as frequently due to unclarity in what is being said as to any failings in the framework.

Second, this work is perhaps unusual in largely eschewing the notion of identity (between phenomenal and physical states, say) in favor of the notion of supervenience. I find that discussion framed in terms of identity generally throws more confusion than light onto the key issues, and often allows the central difficulties to be evaded. By contrast, supervenience seems to provide an ideal framework within which the key issues can be addressed.

The notion of supervenience alone can be a recipe for loose philosophy, as it is often unclear just what "A is supervenient on B" comes to. To tighten things up, we must focus on the *modality* of the supervenience connection; is it underwritten by logical necessity, mere nomic necessity, or something else? It is widely agreed that consciousness supervenes on the physical in some sense; the real question is how tight the connection is. Modality is the key issue here. Discussions that ignore this issue generally avoid the hardest questions about consciousness; I think this is responsible for the unsatisfying nature of much published work on consciousness. Those skeptical of modal notions will be skeptical of my entire discussion, but I think there is no other satisfactory way to frame the issues.

The final word on consciousness will not be said for a long time. This work is at best a step in that direction. Consciousness poses such a baffling problem that sometimes it can seem quite beyond our capacities to understand. I hope, minimally, that this work demonstrates the possibility of making progress on the problem of consciousness without denying its existence or reducing it to something it is not. The problem is fascinating, and the future is exciting.

No. Xia stopped, twirling toward him in slow motion. Her icy mint eyes grew wide. *You're in danger here.* Panic whitened her face as she stared toward the house. *Go home now. Before it's too late. And find me the antidote.*

What kind *of antidote?*

Xia disappeared beyond the junipers, yet her final message burst into Joey's mind like the pop of a firecracker: *The antidote for zombie poison.*

Dian Curtis Regan, *My Zombie Valentine.* Scholastic Inc., 1993.

Chapter 1

Two Concepts of Mind

1.1 What is consciousness?

Conscious experience is at once the most familiar thing in the world and the most mysterious. There is nothing we know about more directly than consciousness, but it is far from clear how to reconcile it with everything else we know. Why does it exist? What does it do? How could it possible arise from lumpy grey matter? We know consciousness far more intimately than we know the rest of the world, but our *understanding* of the rest of the world far outstrips our understanding of consciousness.

Consciousness can be startlingly intense. It is the most vivid of phenomena; nothing is more real to us. Nevertheless, it can be frustratingly diaphanous. When it comes to talking about conscious experience, it is notoriously difficult even to pin down the subject matter in a way that facilitates communication. *The International Dictionary of Psychology* does not even try to give a straightforward characterization:

> *Consciousness*: The having of perceptions, thoughts, and feelings; awareness. The term is impossible to define except in terms that are unintelligible without a grasp of what consciousness means. [...] Consciousness is a fascinating but elusive phenomenon: it is impossible to specify what it is, what it does, or why it evolved. Nothing worth reading has been written about it. (Sutherland 1989.)

Almost anyone who has thought hard about consciousness will have some sympathy with these sentiments. Even the limited attempt at a definition here is disputable. It is arguable that there can be perception and thought that is not conscious. What is central to consciousness, at least in the most interesting sense thereof, is *experience*. But this is not definition; at best, it is clarification.

Trying to define conscious experience in terms of more primitive notions is fruitless. One might as well try to define *matter* in terms of something more fundamental. The best we can do is to give various characterizations that lie at the same level. These characterizations cannot qualify as true definitions, due to their implicitly circular nature, but they can help to pin down what is being talked about. I presume that every reader has conscious experiences of their own. If all goes well, these characterizations will help establish that it is just *those* that we are talking about.

Consciousness is perhaps best characterized as 'the subjective quality of experience'. When we perceive, think, and act, there is a whir of causation and information-processing, but this processing does not usually go on in the dark. There is also an internal aspect; there is something it feels like to be a cognitive agent. This internal aspect is conscious experience. Conscious experiences range from vivid color sensation to experiences of the faintest background aromas; from hard-edged pains to the elusive experience of thoughts on the tip of one's tongue; from mundane everyday sounds and smells to the encompassing grandeur of musical experience; from the triviality of a nagging itch to the weight of a deep existential angst; from the specificity of the taste of peppermint to the generality of one's experience of selfhood. All these have a distinct experienced quality; all are prominent parts of the inner life of the mind.

We can say that a being is conscious if there is *something it is like* to be that being, to use a phrase made famous by Nagel (1974). Similarly, a mental state is conscious if there is something it is like to be in that mental state. Equivalently, we can say that a mental state is conscious if it has a *qualitative feel*—an associated quality of experience. These qualitative feels are also known as phenomenal qualities, or *qualia* for short. The problem of explaining these phenomenal qualities is just the problem of explaining consciousness. This is the really hard part of the mind–body problem.

Why should there be conscious experience at all? Although from a subjective viewpoint it is the most familiar element of nature, from an objective viewpoint it is an utterly unexpected feature. Taking the objective viewpoint, we can certainly tell a story about how fields and particles in the spatiotemporal manifold interact in complex ways, leading to the development of complex systems such as brains. In principle, there is no deep philosophical mystery about the fact that these systems can process information in complex ways, react to stimuli with sophisticated behavior, and even exhibit such complex capacities as learning, memory, and language. All this is impressive, but it is not metaphysically baffling. The existence of conscious experience, on the other hand, seems to be a *new* feature from this viewpoint—it is not something that one would have predicted from all the other features alone.

We can call this the *Surprise Principle*: if all we knew about were the facts of physics, and even the facts about causation and information-processing in complex systems, then conscious experience would come as a complete surprise to us. Unless one had had conscious experience oneself, there would be no compelling reason to postulate the phenomenon. The hypothesis would seem unwarranted; almost mystical, perhaps. Yet there is conscious experience; we know about it more directly than we know about anything else. The question is, how do we reconcile conscious experience with everything else we know?

Conscious experience is a natural phenomenon. It is part of the natural world, and like other natural phenomena it cries out for explanation. There are at least two major explananda here. The first and most central is the very *existence* of consciousness. Why does conscious experience exist? If it arises from physical systems, as seems likely, how does it do arise? This leads to some more specific questions. Is consciousness itself physical, or is it merely a concomitant of physical systems? How widespread is consciousness? Do mice, for instance, have conscious experience?

A second explanandum is the *nature* of conscious experiences. Given that conscious experience exists, why do experiences have their specific character? Why is seeing red like *this*, rather than like *that*? It seems coherent that when looking at red things, such as roses, one might have had the sort of color experiences that one in fact has when looking at blue things. Why is the experience one way rather than the other? Why, for that matter, do we experience the reddish sensation that we do, rather than some entirely different kind of sensation, like the sound of a trumpet?

When someone strikes middle C on the piano, a complex chain of events is set into place. Sound vibrates in the air and a wave travels to my ear. The wave is processed and analyzed into frequencies inside the ear, and a signal is sent to the auditory cortex. Here there is further processing: isolation of certain aspects of the signal, categorization, and ultimately reaction. All this is not so hard to understand in principle. But why should this be accompanied by an *experience*? And why, in particular, should it be accompanied by *that* experience, with the characteristic rich tone and timbre that we associated with middle C on a piano? It is these two questions that we would like a theory of consciousness to answer.

Ultimately one would like a theory of consciousness to do at least the following: (a) give the conditions under which physical processes give rise to consciousness; and (b) for those processes that give rise to consciousness, specify just what sort of experience will be associated. All this, of course, should ideally be done by a theory that makes the arising of consciousness intelligible, without any appeals to magic; one can hope that consciousness will be seen as an integral part of the natural world. Currently it may be hard to see the form that such a theory might take, but without such a theory we could hardly be said to fully understand consciousness.

Before proceeding, I should note that the term 'consciousness' has some other senses.[1] Sometimes it is used to refer to a cognitive capacity, such as the ability to introspect or to report one's mental states. Sometimes it is used synonymously with 'awakeness'. Sometimes it is meant in a way closely tied to our ability to focus attention, or to voluntarily control our behavior. Sometimes 'to be conscious of something' comes to the same thing as 'to know about something'. All of these are valid senses of the term, but all pick out phenomena distinct from what I am talking about, and phenomena that are significantly less difficult to explain. I will say more about these alternative notions of consciousness in section 1.5, but for now, when I talk about consciousness, I will be talking only about the subjective quality of experience: the way it *feels* to be a cognitive agent. (The expression 'conscious experience' is less ambiguous than 'consciousness' alone, even if it is something of a tautology, so I will use it frequently.)

A catalog of conscious experience

Conscious experience can be fascinating to attend to. Experience comes in an enormous number of varieties, consisting in many different aspects. A far-from-complete catalog of the aspects of conscious experience might include those in the following pre-theoretical, impressionistic list. Nothing here should be taken too seriously as philosophy, but it should help to focus attention on the subject matter at hand.

Visual experiences. Among the many varieties of visual experience, color sensations stand out as perhaps the paradigm examples of conscious experience, due to their pure, seemingly ineffable qualitative nature. Some color experiences can seem particular striking, and so can be particularly good as focusing our attention on the mystery of consciousness. In my environment now, there is a shade of deep purple from a book upon my shelf that is particularly rich; an almost surreal shade of green in a photograph of ferns near a waterfall in a poster on my wall; and a sparkling array of bright red, green, orange, and blue lights on a Christmas tree that I can see through my window. But any color can be awe-provoking if we attend to it, and reflect upon its nature. Why should it feel like *that*? Why should it feel like anything at all? How could I possibly convey the nature of this color experience to someone who had not had such an experience?

Other aspects of visual experience include the experience of shape, of size, of brightness and of darkness. A particularly subtle aspect is our experience of depth. As a child, one of my eyes had excellent vision, but the other was very poor. Because of my one good eye, the world looked crisp and sharp, and it certainly seemed three-dimensional. One day I was fitted with glasses, and the change was remarkable. The world was not much sharper than before, but it suddenly looked *more* three dimensional: things that had depth before somehow got deeper, and the world seemed a richer place. If you cover one eye and then uncover it, you can get an idea of the change, though in a watered-down form. In my previous state, I would have said that

[1]There are also many terms that pick out more or less the *same* class of phenomena as 'consciousness' in its central sense. These include: 'experience', 'qualia', 'phenomenology', 'phenomenal', 'subjective experience', and 'what it is like to be'. Apart from grammatical differences, the differences between these terms are most subtle matters of connotation. 'To be conscious' is roughly synonymous with 'to have qualia', or 'to have subjective experience', and so on. Any differences in the class of phenomena picked out are insignificant. I will use the terms interchangeably up to grammar. (Like 'consciousness', many of these terms (e.g., 'experience' and 'phenomenology') are somewhat ambiguous, but I will never use these terms in their alternative senses.)

there was no way for the depth of my vision to improve; the world already seemed as three-dimensional as it could be. The change was subtle, almost ineffable, but extremely striking. Certainly there is an intellectual story one can tell about how binocular vision allows information from each eye to be consolidated into information about distances, thus enabling more sophisticated control of action, but somehow this causal story does not say anything about the way the experience *felt*. Why that change in processing should be accompanied by such a remaking of my experience was mysterious to me as a ten-year-old, and is still mysterious today.

Auditory experience. Sounds, in some ways, are even stranger than visual images. The structure of images usually corresponds to the structure of the world in some straightforward way, but sounds can seem quite independent. My telephone receives an incoming call; an internal device vibrates; a complex wave is set up in the air, eventually reaching my eardrum; and somehow, almost magically, I hear a *ring*. Nothing about the quality of the ring seems to correspond directly to any structure in the world, although I certainly know that it originated with the speaker, and that it is determined by a waveform. But why should that waveform, or even these neural firings, have given rise to a sound quality like *that*?

Musical experience is perhaps the richest aspect of our auditory experience, although our experience of speech must be close. Music is capable of washing over and completely absorbing us, surrounding us in a way that our visual field can surround us, but in which auditory experiences usually do not. One can analyze aspects of our musical experience by breaking the sounds we perceive into notes and tones with complex interrelationships, but the experience of music somehow goes beyond this. A unified qualitative experience arises from a chord, but not from randomly selected notes. An old piano and a far-off oboe can combine to produce an unexpectedly haunting experience. As always, when we reflect, we ask the question: why should *that* feel like *this*?

Tactile experiences. Textures provide another of the richest quality spaces that we experience: think of the feel of velvet, and contrast it to the texture of cold metal, or a clammy hand, or a stubbly chin. All of these have their own unique quality. The tactile experience of water, of cotton candy, or of another person's lips are different again.

Olfactory experiences. Think of the musty smell of an old wardrobe, the stench of rotting garbage, the whiff of newly-mown grass, the warm aroma of freshly-baked bread. Smell is in some ways the most mysterious of all the senses, due to the rich, intangible, indescribable nature of smell sensations. Ackermann (1990) calls it "the mute sense; the one without words". While there is something ineffable about any sensation, the other senses have properties that facilitate some description. Visual and auditory experiences have a complex combinatorial structure that can be described. Tactile and taste experiences generally arise from direct contact with some object, and a rich descriptive vocabulary has been built up by reference to these objects. Smell has little in the way of apparent structure, and often floats free of any apparent object, remaining a primitive presence in our sensory manifold. The primitiveness is perhaps partly due to the slot-and-key process by which our olfactory receptors are sensitive to various kinds of molecules. It seems arbitrary that a given sort of molecule should give rise to *this* sort of sensation, but give rise it does.

Gustatory experiences. Psychophysical investigations tell us that there are only four independent dimensions of taste-perception: sweet, sour, bitter, and salt. But this four-dimensional space combines with our sense of smell to produce a great variety of possible experiences: the taste of Turkish Delight chocolate, of curried black-eye pea salad,[2] of a fillet steak, of a peppermint Lifesaver, of a ripe peach.

Experiences of hot and cold. An oppressively hot, humid day and a frosty winter's day produce strikingly different qualitative experiences. Think also of the heat sensations on one's skin from being close to a fire, and the hot-cold sensation that one gets from touching ultra-cold ice.

Pain. Pain is another paradigm example of a conscious experience, beloved by many philosophers. Perhaps this is because pains form a very distinctive class of qualitative experiences, and are difficult to map directly onto any structure in the world or in the body, although they are usually associated with some part of the body. There are a great variety of pain experiences, from shooting pains and fierce burns through sharp pricks to dull aches.

Other bodily sensations. Pains are only the most salient kind of sensations associated with particular parts of the body. Others include headaches (which are perhaps a class of pain), hunger pangs, itches, tickles, and the experience associated with the need to urinate. Many bodily sensations have an entirely unique quality, different in kind from anything else in our experience: think of orgasms, or the feeling you get when you hit your funny bone. There are also experiences associated with proprioception, our experience of where our body is in space.

---

[2]Cook 12 cups of dried black-eyed peas in boiling water to which 4 tablespoons of salt have been added. Cook until tender, and immerse in cold water. Combine 2 diced red peppers, 5 diced green peppers, 2 diced large onions, 3 cups of raisins and a bunch of chopped cilantro in a dressing made of 1.5 cups of corn oil, 0.75 cups of wine vinegar, 4 tablespoons of sugar, 1 tablespoon of salt, 4 tablespoons of black pepper, 5 tablespoons of curry powder, and a half tablespoon of ground cloves. Serve chilled. Thanks to Lisa Thomas and the Encore Café.

Mental imagery. Moving now toward experiences that are not associated with any particular object in the environment or the body, but that are in some sense internally generated, mental images are a good example. There is often a rich phenomenology associated with visual images conjured up in one's imagination, though not nearly as detailed as those derived from direct visual perception. There are also the interesting colored patterns that one gets when one closes one's eyes and squints, and the strong after-images that one gets after looking at something bright. One can have similar kinds of auditory "images" conjured up by one's imagination, and even perhaps tactile, olfactory, and gustatory images, but these are hard to pin down and any associated qualitative feel is faint at best.

Conscious thought. Many of the things we think and believe do not have any particular qualitative feel associated with them, but many do. This applies particularly to explicit, occurrent thoughts that one thinks to oneself, and to various thoughts that affect one's stream of consciousness. It is often extremely hard to pin down just what the qualitative feel of an occurrent thought is, but it is certainly there. There is *something* it is like to be having such thoughts.

When I think of a lion, for instance, there seems to be just a whiff of some leonine quality to my phenomenology, though this is faint at best. More obviously, cognitive attitudes such as desire often have a strong phenomenal flavor. Desire seems to exert a phenomenological "tug" on us; memory often has a qualitative component, such as the experience of nostalgia or regret, and so on.

Emotions. Emotions often have quite specific qualitative feels associated with them. The sparkle of a happy mood, the weariness of a deep depression, the all-encompassing red-hot glow of a rush of anger, the melancholy of regret; all of these can affect our conscious experience in a profound way, although in a much less specific way than easily-isolatable experiences such as sensations. These emotions pervade and color all of our conscious experiences while they last.

There are other sorts of more transient feelings that lie partway between emotions and the more obviously cognitive aspects of mind. The rush of pleasure one feels when one gets a joke is one example. Another is the feeling of tension one gets when watching a suspense movie, or when waiting for an important event. The butterflies in one's stomach that sometimes accompany nervousness provide a third example.

The sense of self. One sometimes feels that there is something to conscious experience that transcends all these specific elements: some kind of background hum, for instance, that is somehow fundamental to consciousness and that is there even when the other components are not. Perhaps this is an illusion, and our consciousness is merely the sum of specific elements like those listed, but there seems to be *something* to this intuition, even it is very hard to pin down.

This catalog covers a number of bases, but leaves out as much as it puts in. I have said nothing, for instance, about dreams, arousal and fatigue, intoxication, or the novel character of other drug-induced experiences. There are also many rich experiences that derive their character from the combination of two or many of the components from above. We have already considered the combined effects of smell and taste, but an equally salient example might be the combined experience of music and emotion, which interact with each other in a subtle, difficult-to-separate way. I have also said nothing about the unity of conscious experience—the way that all of these experiences seem to be tied together as the experience of a single experiencer. Like the sense of self, it is possible that this unity may be some kind of illusion—it is certainly harder to pin down than any specific experiences—but there is a strong intuition that some unity is there.

Sad to say, we will not be involved this closely again with the rich varieties of conscious experience. In addressing the deep philosophical mysteries associated with conscious experience, a simple color sensation raises the problems as deeply as one's experience of a Bach chorale. The deep issues cut across these varieties in a way that renders consideration of the nature of specific experiences not especially relevant. Still, I hope that this brief look at the rich varieties of conscious experience has helped focus attention on just what it is that we are talking about, and has provided a stock of examples that can be kept in mind during more abstract discussion.[3]

1.2 The phenomenal and the psychological concepts of mind

Conscious experience is not all there is to mind. To see this, observe that although modern cognitive science has had almost nothing to say about consciousness, it has had much to say about mind in general. The aspects of mind with which cognitive science is concerned are different. Cognitive science deals in the explanation of behavior, and insofar as it is concerned with mind at all, it is concerned with mind construed as the internal basis of behavior, and with mental states construed as those states relevant to the causation and explanation of behavior. Such states may or may not be conscious. From the point of view of cognitive science, an internal state responsible for the causation of behavior is equally mental whether it is conscious or not.

At the root of all this lie two quite distinct concepts of mind. The first is what we

[3]For a wealth of reflection on the varieties of specific experiences, see Ackerman's *A Natural History of the Senses* (1990), which provides material for those absorbed by their conscious experience to mull over for days.

may call the *phenomenal* concept of mind. This is the concept of mind as conscious experience, and of a mental state as a consciously experienced mental state. This is the aspect of mind which is most perplexing and with which we will be most concerned, but it does not exhaust the mental.

The second is the *psychological* concept of mind. This is the concept of mind as the causal or explanatory basis for behavior. A state is mental in this sense if it plays a certain kind of causal role in the production of behavior, or at least if plays some appropriate role in the explanation of behavior. On the psychological concept, it matters little whether a mental state has a conscious feel or not. What matters is the role it plays in our cognitive economy.

We might say that on the phenomenal concept, mind is characterized by the way it *feels*, and on the psychological concept, mind is characterized by what it *does*. There should be no question of a competition between these two notions of mind. Neither of them is the "correct" analysis of mind. They are dealing with different phenomena, both of which are quite real. It is only the fact that both are called 'mind' that gives the appearance of competition. In future, I will try to avoid this appearance by using the terms 'phenomenal' and 'psychological' explicitly, and using 'mind' as a coverall.

I will sometimes speak of the phenomenal and psychological 'aspects' of mind, and sometimes of the 'phenomenal mind' and the 'psychological mind'. At this early stage, I do not wish to beg any questions about whether the phenomenal and the psychological will "turn out" to be the same thing. Perhaps every phenomenal state will be a psychological state, in that it plays some role in the causation and explanation of behavior; and perhaps every psychological state will turn out to have some intimate relation to the phenomenal. For now, all that counts is the *conceptual* distinction, or the distinction in intension between the two notions. What it *means* for a state to be phenomenal is for it to feel a certain way, and what it means for a state to be psychological is for it to play an appropriate causal role. These are quite distinct notions that should not be conflated, at least at the outset.

Within both the phenomenal and the psychological concepts of *mind*, there are many more specific mental concepts that fall into one or the other domains. Any given mental state-type (or mental property) can be analyzed one way or the other. For instance, sensation, in its central sense, is best taken as a phenomenal state-type: what it is for something to be a sensation is for it to have a certain sort of feel. On the other hand, something like learning or memory is best taken as a psychological notion. For something to learn, at a first approximation, is for it to adapt its behavioral capacities appropriately in response to certain kinds of environmental stimulation. In general, a phenomenal state-type is characterized by the way it feels, while a psychological state-type is characterized by its role in the causation and/or explanation of behavior.[4]

Of our mental concepts, perhaps the majority are psychological notions. Such intentional notions as belief and desire are probably best regarded as being defined by their causal role. While some beliefs and desires may have have an associated phenomenal quality, there is little reason to think that the quality is essential to that state's being a belief or a desire: if the state had lacked any phenomenal quality, but played a similar causal role, we would probably still call it a belief. (This is somewhat controversial; I will discuss it more later.) Other mental notions such as action, attention, and even perception are similarly best taken as psychological.

Many mental terms straddle the fence to some degree, and can be taken in both a phenomenal and a psychological sense. Pain, for example, can be defined either as a particular sort of unpleasant phenomenal quality, or as a type of state that tends to be produced by tissue damage and that leads to aversion reactions. (Alternatively, one might say that there is a single sense of 'pain', defined to some extent in terms of both phenomenal and psychological properties. Little rests on this semantic decision.) Even 'sensation' can be seen to have a non-central psychological sense, while perception has a non-central phenomenal sense. This sort of ambiguity can complicate discussion, but it is harmless as long as it is noted explicitly. I will return to this matter later.

A potted history

The phenomenal and the psychological aspects of mind have a long history of being conflated. Descartes was perhaps responsible for getting this conflation underway. With his notorious doctrine that the mind is transparent to itself, he came very close to identifying the mental with the phenomenal. For Descartes, every event in the mind was a *cogitatio*, or a content of experience. To this class he assimilated volitions, intentions, and every type of thought. In his reply to the Fourth Set of Objections, he wrote:

> As to the fact that there can be nothing in the mind, in so far as it is a thinking thing, of which it is not aware, this seems to me to be self-evident. For there is nothing that we can understand to be in the mind, regarded in this way, that is not a thought or dependent on a thought. If it were not a thought nor dependent on a thought it would not belong to

[4]Throughout this discussion, phenomenal and psychological properties and state-types are individuated intensionally, according to the concepts involved. It is possible that in some sense the two different intensions will pick out the same property in extension, as with concepts like 'water' and 'H_2O'. I will discuss the relevance of this later. For now, what matters is the prior intension associated with the concepts.

> the mind *qua* thinking thing; and we cannot have any thought of which we are not aware at the very moment it is in us.

If Descartes did not actually identify the psychological with the phenomenal, he at least assumed that everything psychological that is worthy of being called mental had a phenomenal aspect. The notion of an unconscious mental state, to Descartes, would have been very close to a contradiction in terms.

It was progress in psychological theory rather than in philosophy that was responsible for drawing the two aspects of mind apart. Late 19th-century psychological work, such as that of Wundt and James, can be recognized as a distant descendent of Descartes' tradition in that it looked for the causes of behavior by introspection, and developed psychological theories on the basis of introspective evidence. In this fashion, phenomenology was made the arbiter of psychology. But developments soon after established the psychological as an autonomous domain.

Most notably, Freudian theory established the idea that many important activities of the mind are unconscious, and that there can even be such things as unconscious beliefs and desires. The very fact that this notion seemed coherent is evidence that a non-phenomenal analysis of thought was being used. Rather, the notions were being construed *causally*. Desire, very roughly, was implicitly construed as the sort of state that gives rise to a certain kind of behavior associated with the object of the desire. Belief was construed according to its causal role in a similar way. Of course Freud did not make these analyses explicitly, but something like this was clearly underlying his use of the notions. Explicitly, he recognized that accessibility to consciousness was not essential to a state's relevance in the explanation of behavior, and that a conscious quality was not constitutive of something's being a belief or a desire. Here we can see a notion of a psychological state being developed that is quite independent of phenomenal notions.

Around the same time, the behaviorist movement in psychology had thoroughly rejected the introspectionist tradition. A new brand of psychological explanation was developed that was thoroughly "objective"; there was no room for consciousness in these explanations. Whatever the success of this mode of explanation, it established the idea that psychological explanation could proceed while the phenomenal was ignored. Behaviorists differed in their theoretical positions: some recognized the existence of consciousness but found it irrelevant to psychological explanation, and some denied its existence altogether. Many went further, and denied the existence of *any* kind of mental state, whether phenomenal or psychological. The official reason for this was that internal states were supposed to be methodologically irrelevant in the explanation of behavior, which could be carried out entirely in external terms; but perhaps a deeper reason was that all mental notions were tainted with the disreputable odor of the phenomenal.

In any case, these two developments established as orthodoxy the idea that explanation of behavior is in no way dependent on phenomenal notions. The move from behaviorism to computational cognitive science preserved this orthodoxy, for the most part. This move brought back a role for internal states, which could even be called 'mental' states, but there was nothing particularly phenomenal about them. These states were admissible precisely on the grounds of their relevance in the explanation of behavior. Any associated phenomenal quality was at best beside the point. The concept of the mental as psychological thus had center stage.

In philosophy, the shift in emphasis from the phenomenal to the psychological was codified by Ryle (1949), who argued that all our mental concepts can be analyzed in terms of certain kinds of associated behavior, or in terms of dispositions to behave in certain ways.[5] This view, *logical behaviorism*, is recognizably the antecedent of much of what passes for orthodoxy in contemporary philosophy of psychology. In particular, it was the most explicit codification of the causal/explanatory nature of mental concepts.

Ryle did not put this forward as an analysis of just *some* mental concepts, however. He intended all mental concepts to fall under its grasp. It seemed to many people, as it seems to me, that this view is a non-starter as an analysis of our phenomenal concepts, such as sensation and consciousness itself. To many, it seemed clear that when we talk about phenomenal states, we are certainly not talking about our behavior, or about any behavioral disposition. But in any case, Ryle's analysis provided a promising approach to many other mental notions, such as believing, enjoying, wanting, and pretending, for instance.

Apart from its problems with phenomenal states, Ryle's view had some technical problems. Firstly, it was unclear how behavior or even a behavioral disposition (as opposed to an internal state) could function as a cause of behavior, as might seem appropriate for a mental state. Secondly and more importantly, it was argued (by Chisholm 1957 and Geach 1957) that no mental state-type could be defined by a single range of behavioral dispositions, independent of any other mental states. For instance, if one has the belief that it is raining, one's behavioral dispositions will vary depending on whether one has the desire to get wet. It is therefore necessary to invoke other mental states in characterizing the behavioral dispositions associated with a given mental state-type.

These problems were finessed by *functionalism*, which was developed by Lewis (1966) and most thoroughly by Armstrong (1968).[6] On this view, a mental state is

[5]At least this is the standard interpretation of Ryle's account. His actual views strike me as more subtle.

[6]There are other forms of functionalism, such as that developed by Putnam (1960). I will not be concerned with these here, as they were put forward as empirical hypotheses rather than as analyses of mental concepts.

characterized by its *causal role*: that is, in terms of the kinds of stimulation that tend to produce it, the kind of behavior it tends to produce, and the way it interacts with other mental states. This view made mental states fully internal and able to stand in the right kind of causal relation to behavior, answering the first objection, and it allowed mental states to be defined in terms of their interaction with each other, answering the second objection (such interdefinition need not be circular, as it can ultimately be cashed out in terms of an overall causal structure that produces behavior in the right way).

On this view, then, our mental concepts can be analyzed *functionally*: in terms of their actual or typical causes and effects. To actually give such an analysis for any given mental concept is highly non-trivial. Armstrong (1968) gives a number of such analyses, but these are incomplete. As an in-principle position, however, this seems to provide a reasonable construal of many of our mental concepts, at least insofar as they play a role in the explanation of behavior. For instance, the notion of learning might be analyzed as the adaptation of one's behavioral capacities in response to environmental stimulation. To take a more complex case, a belief that P might be very roughly analyzed as the sort of state that tends to be produced by P being the case, that leads to behavior that would be appropriate if P were true, that interacts inferentially with other beliefs and desires in a certain sort of way, and so on. There is a lot of room for working out the details, but the overall idea seems on the right track.

Like Ryle, however, Armstrong and Lewis did not put this forward as an analysis of *some* mental concepts. Rather, it was meant as an analysis of all mental concepts. In particular, they argued that the notions of experience, sensation, consciousness, and so on, could be analyzed in this fashion. This assimilation of the phenomenal to the psychological seems to me to be as great an error as Descartes' assimilation of the psychological to the phenomenal. It is simply a false analysis of what it means to be phenomenal. When we wonder whether somebody is having a color experience, we are not wondering whether they are receiving environmental stimulation and processing it in a certain way. We are wondering whether they are *experiencing* a color sensation, and this is a distinct question. It is a conceptually coherent possibility that something could be playing the causal role without there being any associated experience.

To put the point a different way, note that this analysis of phenomenal concepts leaves it entirely unclear why anybody was ever bothered by the problem in the first place.[7] There is no great mystery about how a state might play some causal role, although there are certainly technical problems there for science. What is mysterious is why that state should *feel* like something; why it should have a phenomenal quality.

[7]Nagel (1970) makes this point against Armstrong, although with reference to the problem of other minds.

Why the causal role is played and why the phenomenal quality is present are two entirely different questions. The functionalist analysis denies the distinctness of these questions, and is therefore unsatisfactory.[8]

I will return to this matter later, but for now we can note that while the functionalist account is an unsatisfactory analysis of phenomenal concepts, it provides an excellent analysis of other mental notions, such as learning or belief. No parallel worries come up with these notions. It seems no more mysterious that a system should be able to learn than that a system should be able to adapt its behavior in response to environmental stimulation; indeed, these seem to be more or less the same question. Similarly, when we wonder whether somebody has learned something, it seems reasonable to say that in doing this we are wondering whether they have undergone a change that will give rise to an improved capacity to deal with certain situations in the future. Of course a thorough analysis of the concept of learning will be more subtle than this first approximation, but the further details of the analysis will be spelled out within the same framework.

Indeed, the functionalist account corresponds precisely to the definition I have given of *psychological* properties. Most non-phenomenal mental properties fall into this class, and can therefore be functionally analyzed. There is certainly room for arguing over the details of how the functionalist analyses should be specified—not just details of specific analyses, but such framework questions as whether causation, explanation, or both ought to be the relation providing the defining link between psychological properties and behavior, and also questions concerning the role of the environment in characterizations of these properties, but these details are relatively unimportant for our purposes, as long as we recognize the key point that non-phenomenal mental properties are characterized primarily by their role in our cognitive economy.

The moral of the discussion above is that both the psychological and the phenomenal are real and distinct aspects of mind. At a first approximation, phenomenal concepts deal with the first-person aspects of mind, and psychological concepts deal with the third-person aspects (although this is complicated by the fact that one can introspect psychological states such as beliefs without any phenomenal quality being involved). One's approach to the mind will be quite different depending on what aspects of the mind one is interested in. If one is interested in the mind's role in bringing about behavior, one will focus on psychological properties. If one is interested in the conscious experience of mental states, one will focus on phenomenal properties. Neither the phenomenal nor the psychological should be defined away in terms of the other. Conceivably some deep analysis might reveal a fundamental link between

[8]Those who feel that I am begging the question against the functional analysis of phenomenal notions should await more detailed arguments in Chapter 3. For now, one can simply note the strong *prima facie* case for a distinction.

the phenomenal and and the psychological, but this would be a highly non-trivial task, and is not something to be accomplished by prior stipulation. To assimilate the phenomenal to the psychological prior to some deep explanation would be to trivialize the problem of conscious experience; and to assimilate the psychological to the phenomenal would be to vastly limit the role of the mental in explaining behavior.

1.3 The double life of our mental terms

It seems reasonable to say that every mental property is either a phenomenal property, a psychological property, or some combination of the two. Certainly, if we are concerned with those manifest properties of the mind that cry out for explanation, we find first, the varieties of conscious experience, and second, the causation of behavior. There is no third kind of manifest explanandum, and the first two sources of evidence—experience and behavior—provide no reason to believe in any third kind of non-phenomenal, non-functional properties. There are certainly phenomena such as intentionality—the way in which mental states are *about* things in the world—but it is very plausible that intentional properties can be assimilated to one class or the other, as I will discuss shortly. In general, when we leave aside the phenomenal aspects of mind, causal analyses seem quite adequate.

Things are complicated by the already-noted fact that many of our everyday mental concepts have both a psychological and a phenomenal component. For instance, there is a sense in which 'perception' implies the conscious experience of what is being perceived, as well as implying a certain sensitivity and reactivity to environmental stimuli. The possibility of subliminal perception counts against such a notion, but some might argue this only qualifies as perception in a different sense of the term. In any case, it is clear enough why phenomenal and psychological properties tend to be run together like this: it is because the relevant properties tend to co-occur. Generally, when one is receiving stimulation from the environment and processing it in a certain way, some kind of phenomenal quality is instantiated. This is not a conceptual truth—to instantiate a phenomenal quality does not *mean* to process stimulation in a certain way—but it is a fact about the way the world is. It is a commonplace that our categories tend to bind together properties that occur together: witness Wittgenstein's discussion of the category 'game'.

There is little point trying to legislate our everyday mental concepts to one side or the other. Nothing important rests on the question of whether some phenomenal quality is *really* essential for something to count as 'perception' or not. Instead, we can recognize the two different components associated with a concept and explicitly distinguish them, talking for example of 'psychological perception' and 'phenomenal perception'. Our everyday concept of perception presumably combines these two in some subtle weighted combination, but for philosophical discussion, things are much clearer if we keep them separate.

That being said, it is certainly the case that some everyday categories lean strongly toward the phenomenal, and some lean toward the psychological. 'Sensation' is an example of the first, and 'perception' is an example of the second, as is witnessed by the fact that unconscious perception seems to make more sense than unconscious sensation. When I use the term 'perception' alone, it should therefore be taken to refer to the psychological property, without any phenomenal component, whereas 'sensation' will be used for the phenomenal notion unless otherwise specified. A good test for whether a mental notion M is primarily a psychological notion is to ask oneself: could something be an instance of M without any particular associated phenomenal quality? If so, then M is probably a psychological type.

I will go through a number of important everyday mental concepts, examining their phenomenal and psychological components and determining which aspect is most central.

Perception. As we have just seen, 'perception' appears to be mostly a psychological notion, referring to a process whereby environmental stimulation is received and processed in a certain sort of way, playing an appropriate role in our cognitive economy and being available to ultimately influence those aspects of our behavior that related to the object of the perception. All the details are covered up by phrases like "a certain sort of way" and "appropriate role", of course. We will not be concerned with those details, which are a fascinating subject for philosophical analyses in their own right. (See Armstrong 1968 for approximate functional analyses of many mental notions, and Dretske 1970 for an analysis of perception in particular.)

There is perhaps a phenomenal sense of 'perception' which requires the conscious experience of what is being perceived, but this is non-central. Certainly the many cognitive scientists studying perception are studying its causal role, rather than its phenomenal quality. When I use the term 'perception' alone from now on, it will be the psychological sense that I have in mind.

Sensation. This term is the phenomenal twin of 'perception', referring primarily to a phenomenal quality while perhaps having a non-central psychological sense as well. The term 'sensation' is sometimes used to refer to the reception and peripheral processing of certain kinds of stimulation, independent of whether it has any phenomenal quality. Certainly biologists will speak of a sea-slug "sensing" the world, without making any commitment about existence of a conscious experience. Perhaps they would be less inclined to use the word 'sensation', but one sees it occasionally.

Once again, little rests on the decision about how the term should be used, as long we keep the distinctions clear. For our purposes, it is the phenomenal sense of 'sensation' that is central, so when the term is used alone, the phenomenal quality is what is being referred to.

Pain. This is an interesting example of a term that straddles the boundary between the phenomenal and the psychological. Perhaps the most salient aspect of pain is its unpleasant phenomenal quality, but there is also a clear functional notion that goes along with this, where 'pain' is construed roughly as the kind of state that tends to be produced by tissue damage and that leads to withdrawal reactions, changes in the focus of attention, exclamations, and so on.

One can tie oneself into all kinds of philosophical knots by worrying about whether the phenomenal quality or the functional role is more essential to pain. For instance, would a hypothetical system in which all the functional criteria were satisfied but in which the conscious experience were not present be truly in pain? One might be tempted to say no, but this decision might be affected by the observation that we speak about pains that last for a day, even though there are times when they are not conscious. In any case, our answer to this question does not reflect any fact about pain itself. It is simply a semantic decision, in the pejorative sense. I will not legislate any answer to this question, instead speaking of 'pain sensations' and 'psychological pain' when clarification is necessary.

Learning, memory, categorization. These are almost entirely psychological concepts. To learn is just for one's cognitive capacities to adapt in a certain way. To remember is just to have internal access to information about the past, at a first approximation. To categorize is just to divide up objects and representations in a certain sort of way. Any faint phenomenal tinges that these concepts might have are derivative on their link to the notion of belief, which I will discuss below. Overall, however, the presence or absence of a phenomenal quality makes little difference to whether some process counts as learning or as categorization. Perhaps the phenomenal connotations of 'memory' are stronger, but in any case I will use all of these as paradigm psychological notions.

Belief. The status of belief is somewhat less clear, mostly because the functional criteria for belief are much less straightforward to spell out than those for perception, learning, and the like. Overall, however, it is generally accepted that belief is best analyzed in terms of its role in our cognitive economy and its link to behavior, no matter what the complications are with the details.

The most problematic aspect of belief is its semantic aspect, or its intentionality: the fact that a belief is *about* something in the world. Analyzing just what this "aboutness" comes to poses significant philosophical problems. Some have even seen this as posing a mystery of the same kind that phenomenal properties pose. On this view, intentionality is taken to be unanalyzable in terms of more primitive notions.

There is a clear disanalogy between intentionality and conscious experience, however. The former is something of a *explanatory construct.* We attribute beliefs to ourselves and others largely on the basis of associated causal processes, and in particular according to their success in explaining behavior. Conscious experience, by contrast, is no explanatory construct. Rather, it is a manifest *explanandum* in its own right. Phenomenal qualities force themselves on us in a direct way that intentional properties do not. Our awareness of intentional properties is relatively indirect, and it is this indirectness that allows the notion of intentionality to be explicated in terms of other properties.[9]

It may be the case that a causal analysis of intentionality leaves something out, but what is left out seems only to be our *experience* of intentionality—that is, intentionality's phenomenal aspect. It follows that intentionality does not pose any third kind of problem, but is subsumed by the other two kinds. Explaining our experience of intentionality—for instance, the fact that my conscious experience of a tree seems to be *about* a tree—is a difficult matter, but it is part of the problem of consciousness rather than something new.

Some philosophers (e.g. Searle 1990, Goldman 1993) have argued that a phenomenal quality is *constitutive* of intentional content, but this has not met with much agreement. It is certainly true that many intentional states have some conscious feel. Perception is an obvious example, but there is also a phenomenal quality to many of our beliefs and desires, at least insofar as they are occurrent components of our stream of thought. However, there is little reason to hold that this phenomenal quality is part of what *makes* those states beliefs, or what gives them their content. When I have a belief, even an occurrent belief, that Don Bradman was the greatest cricket player of all time, any associated phenomenal quality is elusive at best, and certainly does not seem to play any role in *making* that state a belief about Bradman. It seems reasonable to say I would still have had the belief about Bradman if the phenomenal quality had been quite different, or if there had been none at all. What makes the state a belief about Bradman is its causal connection to Bradman and the role that it plays in my cognitive economy.

A weaker position might hold that there is no *particular* phenomenal quality that is required for a particular belief, but that *some* conscious experience is required for a system to have beliefs at all. Some people have the intuition that for a system without conscious experience to have beliefs would be a conceptual impossibility. This move is somewhat ad hoc, and implies that the role of phenomenal properties in fixing belief

[9]It is also this indirectness that allows the possibility of *eliminativism* about intentional states, of the kind put forward by Churchland (1981). Given that such states are defined in terms of some causal or explanatory role, it is possible that no state will turn out to play such a role, although behavioral evidence may give us grounds to find that extremely unlikely. Our direct experience of consciousness makes eliminativism about consciousness an entirely different kettle of fish. It is one thing to deny the existence of an explanatory construct; it is another to deny the existence of an explanandum.

content is extremely thin. Once we have gone this far, it is more natural to suggest that intentional properties are not conceptually dependent on phenomenal properties.

In any case, as usual we need not legislate this matter. We can note that belief has a phenomenal aspect and a psychological aspect, and nothing important will rest on the question of whether the phenomenal aspect is essential for something to be a *real* belief. If someone insists that it is, we can simply talk about the purely psychological notion 'schmelief', which will end up doing almost all the explanatory work that belief usually does for us. However, insofar as I will be explicitly concerned with belief, I will be concerned with the role it plays in the causation of behavior, so it is more convenient to take the term 'belief' in the psychological sense, recognizing that this captures everything about belief except for its phenomenal aspect. (I will be careful not to beg any important questions by this strategy.)

A similar account can be given of the conceptual status of other intentional properties, such as desiring, hoping, fearing, and so on. The states in question may have some phenomenal aspect, but in each case the causal role is more central to their qualifying as instances of the intentional property in question. It is easiest to assume that these are psychological notions. When we wish to focus on any associated phenomenal aspects, we can do so explicitly.

Most intentional properties depend on one's environment as well as on one's internal processes. Arguments by Putnam (1975) and Burge (1979) establish that the content of a belief depends on external factors. More obviously, for a state to qualify as knowing-that-P, the requirement that P be true poses a strong external constraint. We can handle this on the psychological account in two ways. Firstly, we might ensure that the causal roles invoked in analyses of these properties stretch into the environment, requiring for example that a belief about water have a causal link with water itself, and that knowledge about P be somehow caused by P being the case. Alternatively, we might divide these properties into an internal and an external part, analyzing the former in terms of some internal causal role in the production of behavior, and leaving the latter as some further relational criterion. The former has the advantage of allowing these properties to cleanly fit our definition of psychological properties. The latter would force the definition to be extended, but might capture some people's intuition that external aspects are irrelevant to the causation and explanation of behavior (see, e.g., Fodor 1987). Even if we made the second choice, little that is important to the distinction between the phenomenal and the psychological would change.

Emotions. Emotions, more clearly than beliefs, have a phenomenal aspect. When we think of happiness and sadness, a distinct variety of conscious experience comes to mind. Once again, however, it is not obvious whether the phenomenal aspect is essential to a state's being an emotion. There is clearly a strong associated psychological property as well. As usual, we need not make any decision on this matter. We can simply talk about the psychological and phenomenal aspects of emotion, and observe that these exhaust the aspects of emotion that require explanation. There is the way an emotion feels, and what an emotion does, and that is all.

The co-occurrence of phenomenal and psychological properties

It is a fact about the human mind that whenever some phenomenal property is instantiated, some corresponding psychological property is instantiated. Conscious experience does not go on in a vacuum. It is always tied to cognitive processing, and it is very likely that in some sense it arises from that processing. Whenever one has a sensation, for instance, there is some information-processing going on: a corresponding perception, if you like. Similarly, whenever one has the conscious experience of happiness, the functional role associated with happiness is generally being played by some internal state. Perhaps it is logically possible that one could have the experience without the causation, but it seems to be an empirical fact that they go together.

In the face of this co-occurrence, the faint-hearted may be tempted to worry whether any real distinction is being made. But it is clear that an intensional distinction is present, no matter whether the extensions coincide. One can wonder how to explain the phenomenal quality, and one can wonder how to explain the playing of the causal role, and these are two distinct wonderings.

That being said, the co-occurrence of phenomenal and psychological properties reflects something deep about our phenomenal concepts. We have no independent language for describing phenomenal qualities. As we have seen, there is something ineffable about them. Although greenness is a very distinct sort of sensation with a rich intrinsic character, there is very little that one can say about it other than that it is green. In talking about phenomenal qualities, we generally have to specify the qualities in question in terms of associated external properties, or in terms of associated causal roles. Our *language* for phenomenal qualities is derivative on our non-phenomenal language. As Ryle said, there are no "neat" sensation words.

If one looks at the catalog of conscious experience that I presented in 1.1, the experiences in question are never described in terms of their intrinsic qualities. Rather, I said things like "the smell of freshly-baked bread", "the patterns one gets when closing one's eyes", and so on. Even a term like 'green sensation' is implicitly defined as something like 'the sort of experience one has when looking at grass, trees, and so on'. In general, insofar as we have communicable phenomenal categories at all, these are defined with respect to either their typical external associations, or more generally with respect to the associated kind of psychological states. For instance, when one speaks of the phenomenal quality of happiness, the term 'happiness' is implicitly defined via some causal role—the state where one judges all to be good, one jumps for joy, and so on. This, perhaps, is one interpretation of Wittgenstein's famous remark, "An inner process stands in need of an outward criterion."

This dependence of phenomenal concepts on causal criteria has led some (including Wittgenstein and Ryle, in some of their moods) to suggest that there is nothing to the meaning of our mental concepts beyond the associated causal criteria. There is a certain plausibility to this: if a phenomenal property is always picked out in terms of a psychological property, why not suppose that there is only one property involved? But this temptation should be resisted. When we talk of a green sensation, we are not simply talking about 'the kind of state that is caused by grass and trees'. We are talking about 'the *phenomenal quality* that generally occurs when a state is caused by grass and trees', or, at best, 'the kind of *phenomenal* state that is caused by grass and trees'.[10] It is this residual phenomenal element in the analysis that prevents an analysis in functional terms.

In general, when a phenomenal property is picked out with the aid of a psychological property P, the phenomenal notion is not just 'P'. It is 'the conscious experience that tends to accompany P'. And importantly, the very notion of 'phenomenal quality' or 'conscious experience' is not defined in psychological terms. Rather, the notion of conscious experience, as we saw earlier, is something of a primitive. *If* there were a functional analysis of the notion of 'experience' or 'phenomenal quality', then the analyses in question would yield functional analyses of specific phenomenal properties, but in the absence of such an analysis we cannot suppose any such thing.

We cannot identify the notion 'phenomenal P' with that of 'psychological P' for all the usual reasons: there are two quite distinct intensions there, as witnessed by the fact that there are two distinct explananda. Although 'phenomenal P' is picked out as 'the experience that tends to accompany psychological P', we can coherently imagine

[10] This is a 'topic-neutral' analysis of specific phenomenal notions not unlike those advocated by Place 1956 and Smart 1959. To be an orange experience, very roughly, is to be the kind of experience that is generally caused by oranges. (Place: "...when we describe the after-image as green...we are saying that we are having the sort of experience which we normally have when, and which we have learned to describe as, looking at a green patch of light" (p. 49). Smart: "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"' (p. 150).) But because of the occurrence of the unanalyzed notion of 'experience', this analysis is not sufficient to immediately establish an identification between phenomenal and physical states, in the way that Place and Smart suggested. Smart's account avoids this problem by leaving 'experience' out of the analysis in favor of the equivocal phrase 'something going on'. If 'something going on' is construed broadly enough to cover any sort of state, then the analysis simply seems faulty; if it is construed narrowly as 'some sort of experience', the analysis is closer to the mark but it does not suffice for the conclusion.

a situation in which phenomenal P occurs without psychological P, and vice versa.[11] A Rolls-Royce icon can be roughly analyzed as the kind of icon that is generally found on Rolls-Royce cars, but this does not mean that to be a Rolls-Royce icon is to be a Rolls-Royce car.

This fact gives us some insight into the relative sparseness of our specifically phenomenal vocabulary as opposed to our psychological vocabulary, and it also helps us understand why phenomenal and psychological properties have so often been conflated. For most everyday purposes this conflation does not matter: when one claims that someone is happy, one need not be talking specifically about either the phenomenal quality or the functional role, as they usually go together. However, for philosophical purposes and in particular for the purposes of explanation, to conflate these properties is fatal. It can be tempting, as collapsing the distinction makes the problem of explaining conscious experience suddenly very straightforward; but it is utterly unsatisfactory for precisely that reason. The problem of consciousness cannot be spirited away on purely verbal grounds.

1.4 The two mind–body problems

The division of mental properties into phenomenal and psychological properties has the effect of dividing the mind–body problem into two: an easy part and a hard part. The psychological aspects of mind pose many technical problems for cognitive science, and a number of interesting puzzles for philosophical analysis, but they pose no deep metaphysical enigma. The question "How could a physical system be the sort of thing that could *learn*, or that could *remember*?" does not have the same bite as the corresponding question about sensations, or about consciousness in general. The reason for this is clear. By our analysis above, these properties are functional properties, characterized by their causal roles, so the question "How could a physical system have psychological property P?" comes to the same thing as "How could a state of a physical system play such-and-such a causal role?". This is a question for

[11] I leave aside here the question of how we can group phenomenal properties into types without relying on associated psychological states— that is, the question of what 'an experience like this one' *means*, if we cannot rely on causal criteria of similarity and difference. Is there "brute" similarity and difference between phenomenal properties, such that we can speak of two people instantiating the same phenomenal state? This vexed question is at the core of many of Wittgenstein's discussions, and is also discussed by Shoemaker 1975b and Dennett 1988. It seems extremely plausible that there is indeed a fact of the matter about the sameness and difference of phenomenal experiences. It may be, however, that we will have to leave these sameness relations as primitive. If it turns out that there *is* no fact of the matter about sameness of experiences, then the details of the above would have to be revised, but (contra Dennett) this would not be nearly enough to make the problem of experience go away.

the sciences of physical systems. One simply needs to tell a story about the organization of the physical system that allows it to react to environmental stimulation and produce behavior in the appropriate sorts of ways. While the technical problems are enormous, there is a clearly-defined research program for their answer. The metaphysical problems are relatively few.

This is not to say that psychological properties pose *no* philosophical difficulties. There are significant problems in coming up with the correct analyses of these notions, for instance. Even if it is widely accepted that these are functional concepts, there can be significant disagreement about just how the requisite functional analyses should run. Intentional properties such as belief and desire, for example, provide fertile grounds for argument (Dennett 1987; Fodor 1987). In particular, the question of just what constitutes the content of a given intentional state is still poorly understood (Block 1986; Dretske 1981; Fodor 1987; Millikan 1986). There are also technical problems concerning just how high-level constructs such as these can play a real causal role in the production of behavior (Heil and Mele 1992), especially if these are partly constituted by properties of the environment (Burge 1986), or if there are no strict laws connecting psychological states with behavior (Davidson 1970). Then there are semi-empirical problems in the foundations of cognitive science concerning just how these properties might be instantiated in existing cognitive systems (whether by a language of thought (Fodor 1975) or a connectionist network (Clark 1989)), if indeed they are instantiated at all (Churchland 1981).

These problems are all serious, but they have the character of puzzles rather than mysteries. The situation here is analogous to that in the philosophy of biology, where there is no pressing life–body problem; there are merely a host of technical problems about evolution, selection, adaptation, fitness, and species. Just as the apparent metaphysical mysteries surrounding biology were disposed of long ago, it is fair to say that the mind–body problem for psychological properties is for all intents and purposes dissolved. What remains is a collection of smaller technical problems with which the normal course of scientific and philosophical analysis can grapple.

The phenomenal aspects of mind are a different matter. Here, the mind–body problem is as baffling as it ever was. The impressive progress of the physical and cognitive sciences has not shed significant light on the question of how and why cognitive functioning is accompanied by conscious experience. The progress in the understanding of the mind has almost entirely centered on the explanation of behavior. This progress leaves the question of conscious experience untouched.

If we like, we can view the psychological/phenomenal distinction not so much as splitting the mind–body problem as factoring it into two separate parts. The hardest part of the mind–body problem is the question: how could a physical system give rise to conscious experience? We might factor the link between the physical and conscious experience into two parts: the link between the physical and the psychological, and the link between the psychological and the phenomenal. As we saw above, we now have a pretty good idea of how a physical system can have psychological properties: the *psychological* mind–body problem has been dissolved. What remains is the question of why and how these psychological properties are accompanied by phenomenal properties: why all the stimulation and reaction associated with pain is accompanied by the *experience* of pain, for instance. Following Jackendoff (1987), we can call this residue the *mind–mind problem.* Current physical explanations take us as far as the psychological mind. What remains ill-understood is the link between the psychological mind and the phenomenal mind. (Jackendoff distinguishes the "phenomenological mind" and the "computational mind". This distinction comes to much the same as the phenomenal/psychological distinction outlined here, although I would not like to beg the question about whether psychological properties are computational.)

It is conceivable that the link between the phenomenal and the physical might be independent of that between the psychological and the physical, so that this factoring would be impossible, but it seems unlikely. The close correlation that we have seen between phenomenal and psychological properties suggest a deep link. In later chapters, I will argue that this link is an extremely strong one, and that the factoring strategy is in fact the right approach to the mind–body problem. It follows that understanding the link between the psychological and the phenomenal is crucial to understanding conscious experience.

1.5 Two concepts of consciousness

Any discussion of consciousness is complicated by the fact that the term is ambiguous. After the discussion in 1.3, it will perhaps not be surprising to find out that 'consciousness' has both psychological and phenomenal senses. The phenomenal sense is relatively clear-cut—it is what we have been discussing all along. Indeed, 'consciousness' in this sense is more or less equivalent to 'the existence of some phenomenal quality'. This is the key sense, as it is the one that poses the major explanatory problems. When there is a danger of ambiguity, we can talk about *phenomenal consciousness.*

There are also a number of psychological properties that tend to be subsumed under the term 'consciousness'. For example, the ability to introspect and to report on one's internal states is typically called 'consciousness'. These senses we can group under the category of *psychological consciousness.* This ambiguity can lead to much confusion in discussing consciousness. A frequent occurrence is for someone to claim to be giving an explanation of consciousness, thereby investing the problem with all the gravity of the problem of phenomenal consciousness, but then to give an explanation of some aspect of psychological consciousness, such as our ability to introspect. This explanation might be quite valid in its own right, but one is left with the sense that more has been promised than has been achieved.

Varieties of psychological consciousness

There are numerous psychological notions for which the term 'consciousness' is sometimes used. These include the following.

1. *Awakeness.* Here, 'consciousness' is taken to be the sort of state we are in when awake but not when asleep. That this does not coincide with phenomenal consciousness can be seen from the fact that it is very plausible that we have phenomenal experiences when we are dreaming (although see Dennett 1978d). The relevant functional role might be analyzed in terms of an ability to process information about the world and deal with it in a rational fashion, at a first approximation.

2. *Introspection.* This is the process by which we can become aware of the contents of our internal states. For instance, it is by introspection that I determine that I have a deep-seated loathing for my great-uncle Oswald, or that I realize that I prefer broccoli to pumpkin, or even that I notice that I am hungry. If you ask me about my beliefs, it is by introspection that I determine my answer. This knowledge of one's mental states is an important component of the everyday concept of consciousness, and it is quite clearly a functional notion. One can analyze it in terms of one's rational processes being sensitive to information about one's internal states in the right sort of way, and being able to use this information appropriately.

3. *Reportability.* This closely related notion refers to our ability to report the contents of our mental states. This ability presupposes the ability to introspect, but is more constrained than that ability. It is conceivable that a dog, for instance, might be able to introspect without being able to report the results of its introspection. This notion of consciousness has appealed to some philosophers and psychologists of an operationalistic bent, and to those who wish to tie consciousness closely to language.

4. *Self-consciousness.* This refers to our ability to think about ourselves, our awareness of our existence as individuals and of our distinctness from others. This is something that many psychologists talk about when they discuss "consciousness". My self-consciousness might be analyzed in terms of my access to a self-model, or my possession of a certain sort of representation that is associated in some way with myself. While it is plausible that some degree of phenomenal consciousness is possessed by animals much less sophisticated than ourselves, it may well be that self-consciousness is limited to humans and a few "higher" animals.

5. *Attention.* This is another notion that psychologists talk about in association with consciousness, and is reflected to some degree by everyday usage. In this sense, we are conscious of something when we are paying attention to it: that is, when a significant portion of our cognitive resources are devoted to dealing with some information. We can be phenomenally conscious of something without attending to it. Think of the fringes of the visual field, for instance.

6. *Voluntary control.* In another sense, we say some aspect of *behavior* is conscious when we do it deliberately—that is, with an element of prior thought about what it is that we are going to do.

7. *Knowledge.* In another everyday sense, we say that somebody is conscious of some fact or thing when they simply know that fact or know about that thing. This notion is rarely found in technical writings on consciousness, but it is perhaps as central to the everyday usage of the term as anything else.

That all of these are functional notions can be seen from the answer to the question: how would one explain them? If one were to try to explain attention, one might devise a model of one's cognitive processes that lead to resources being concentrated on one aspect of available information rather than another. If one were to try to explain introspection, one would try to explain the processes by which one is sensitive to one's internal states in the appropriate way. Similar stories apply to explanation of the other properties. In each case, a functional explanation suffices.

It remains the case that there is a phenomenal aspect associated with many or all of these concepts. There is a certain sort of phenomenal quality associated with introspection, and perhaps with self-consciousness. These comprise only a limited range of phenomenal experience, leaving out perceptual experience, among other things, but they are phenomenal nevertheless. As with terms like 'perception' and 'emotion', it is therefore possible for these terms to be used to refer to the phenomenal quality in question. Thus 'self-consciousness' and 'introspection' can both be seen to have phenomenal connotations. In a similar way, there is a phenomenal quality associated with attention, and even with the voluntary control of behavior. One should not confuse the phenomenal property with the psychological property, but as in previous cases, they tend to co-occur.

Consciousness and awareness

We have seen that there is a psychological property associated with our experience of emotion, that there is a psychological property associated with our experience of self-consciousness, that there is a psychological property associated with our experience of sensation, and so on. It might seem plausible to suppose that there is some psychological property associated with experience *per se*, or with phenomenal consciousness, and I think there is such a property. This property is perhaps the most general brand of psychological consciousness.

8. *Awareness.* This might be analyzed roughly as a state wherein we have access to some information, and are able to use that information in the control of behavior. One can be aware of some object in the environment, of a state of one's body, or of some mental state, among other things. Awareness generally brings with it the ability to knowingly direct behavior in a certain way depending on that information. This is clearly a functional notion. On some uses the term 'awareness' is a phenomenal

notion, coming to much the same thing as 'consciousness' itself. I stipulate that I am *not* using the term in the phenomenal sense, but in the functional sense that I have described.

The notion that there is a functional notion of consciousness that can be explicated in terms of access is put forward by Block (1990a), who talks about "phenomenal consciousness" and "access consciousness". His "access consciousness" corresponds closely to the notion of awareness that I am describing. In a similar fashion, Newell (1992) makes an explicit separation between "awareness" and "consciousness". He describes awareness as "the ability of a subject to make its behavior depend on some knowledge", and goes on to spell out the distinction between this notion and consciousness, which he says is a nonfunctional phenomenon.

Many others have suggested a distinction between phenomenal and broadly functional notions of consciousness. Nelkin (1989) distinguishes CN (consciousness in the "Nagel" sense) from C1 (a first-order information-processing state) and C2 (second-order direct non-inferential accessing of other conscious states). Bisiach (1988) distinguishes C1 (phenomenal experience) from C2 (the access of parts or processes of a system to other of its parts or processes). Natsoulas (1978) distinguishes all sorts of senses of the term 'consciousness'. Dennett (1969) distinguishes two kinds of "awareness", the first associated with verbal reports and the second more generally with the control of behavior; neither of these is a phenomenal sense, however.

In general, whenever we have phenomenal consciousness, we seem to have awareness one way or another. My phenomenal experience of the yellow book beside me is accompanied by my functional awareness of the book, and indeed by my awareness of the yellow color. My experience of a pain is accompanied by an awareness of the presence of something nasty, which tends to lead to withdrawal and the like, where possible. The fact that any conscious experience is accompanied by awareness is made clear by the fact that any conscious experience is *reportable*. If we are having some conscious experience, we can talk about the fact that we are having it. We may not be paying attention to it, but we at least have the ability to focus on it and talk about it, if we so choose. This reportability immediately implies that we are aware in the relevant sense. Of course, an animal or a prelinguistic human might have conscious experience without the ability to report. Such a being would plausibly also have a degree of awareness. Awareness does not entail the ability to report, but in a being with sophisticated linguistic abilities, one goes along with the other.[12]

As I have described it, consciousness is always accompanied by awareness, but awareness need not be accompanied by consciousness. One can be aware of some fact without any particular associated phenomenal experience, for instance. However, one

[12]The ability to report does not of course entail the ability to *describe*. We have seen that description of phenomenal experience can be difficult if not impossible.

might be able to put constraints on the notion of awareness so that it turned out to be coextensive with phenomenal consciousness, or nearly so. I will not attempt that project here (but see Chapter 6): spelling it out might involve a particular sort of direct or "high-bandwidth" access as opposed to the low-bandwidth access associated with non-phenomenal awareness, and with a certain capacity to strongly affect our cognitive functioning, where the relevant notions would have to be spelled out appropriately.

The notion of awareness subsumes most or all of the various psychological notions of consciousness enumerated above. Introspection can be analyzed as awareness of some internal state. Attention can be analyzed as a particularly high degree of awareness of an object or event. Self-consciousness can be understood as awareness of oneself. Voluntary control is trickier, although it might be partly analyzed as requiring attention to the behavior one is performing. Awakeness might be roughly characterized as a state in which one is able to deal rationally with one's environment to some extent, and so implies a particular sort of awareness.

Explaining consciousness versus explaining awareness

Awareness, like other psychological properties, poses few metaphysical problems. The problems posed by the psychological varieties of consciousness are of the same order of magnitude as those posed by memory, learning, and belief. Certainly the notion of awareness is not crystal-clear, so there is room for significant philosophical analysis of just what it comes to; and there is room for an enormous amount of research in cognitive science, investigating the ways that real and artificial cognitive systems might function in such a way that they are aware. But the outlines of this research program are reasonably clear, and there is little reason to suppose that the normal course of cognitive science, backed by appropriate philosophical analysis, should not eventually succeed.

Insofar as consciousness is the really difficult problem for a science of the mind, it is phenomenal consciousness that is at issue. The problems here are of an entirely different order of magnitude. Even after we have explained the physical and computational functioning of a cognitive system, we will still need to explain why the system has conscious experiences, if indeed it has any at all. Of course this claim is controversial, and much of the next two chapters will be devoted to justifying it, but for present purposes, we can simply note the difference in the *prima facie* problems that the phenomenal and psychological varieties present us with. It is phenomenal consciousness that is the *worrying* problem.

Given the significant differences between the psychological and phenomenal notions of consciousness, it is unfortunate that they are frequently conflated in the literature. As before, this conflation matters little in everyday speech, as awareness and phenomenal consciousness usually go together. But for the purposes of explanation, the intensional distinction is crucial. Insofar as any remotely satisfactory explanations of "consciousness" have been put forward, it is usually a psychological aspect of consciousness that is explained. The phenomenal aspects generally go untouched.

Many recent philosophical analyses of consciousness have concerned themselves primarily with the non-phenomenal aspects. Rosenthal (1990) argues that consciousness ought to be analyzed as the presence of some sort of higher-order thought—that is, a thought about another mental state. This might be a useful analysis of introspective consciousness, and perhaps of other aspects of awareness, but it tells us little about phenomenal experience. (Rosenthal argues explicitly that the problem of consciousness ought to be separated from the problems associated with sensation. Clearly, we can agree with the separation—phenomenal consciousness and awareness are indeed distinct phenomena—while little rests on the issue of which gets to be called 'consciousness'.)

Dennett (1991) spends much of his book outlining a detailed Pandemonium-style model, which he claims as an explanation of consciousness. On the face of it, what this model provides is an explanation of *reportability*, and perhaps, therefore, of introspective consciousness and even of many aspects of awareness. A more primitive model by Dennett (1978c) addresses the same problem. However, nothing in the model provides any kind of explanation of phenomenal consciousness. (Dennett goes on to argue that there is no further phenomenon to be explained. This view amounts to a kind of eliminativism about phenomenal consciousness; I will discuss it in detail later.)

Armstrong (1968), confronted by consciousness as an obstacle for his functionalist theory of mind, analyzes the notion in terms of the presence of some self-scanning mechanism. This might provide a useful account of self-consciousness and introspective consciousness, but it leaves the problem of phenomenal experience entirely to one side. Armstrong (1981) talks about both perceptual consciousness and introspective consciousness, but is concerned with both only as varieties of awareness, and does not address the problems posed by the phenomenal qualities of experience. Thus the sense in which consciousness is really *problematic* for his functionalist theory is sidestepped, courtesy of the ambiguity in the notion of consciousness.

Others writing on the topic of "consciousness" have been primarily concerned with self-consciousness or introspective consciousness. Van Gulick (1988), in suggesting that consciousness should be analyzed as the possession of "reflexive metapsychological information", is at best providing an analysis of these psychological notions. (He concedes that the phenomenal aspects may be left out by such an analysis.) Similarly, Jaynes' (1976) elaborate theory of consciousness is concerned only with our awareness of our own thoughts. It says nothing about phenomena associated with perception, and therefore could not hope to be a theory of awareness in general, let alone a theory of phenomenal consciousness. Hofstadter (1979) has some interesting things to say about consciousness, but he is more concerned with introspection, free will, and the sense of self than with conscious experience *per se*.

Insofar as consciousness has been a topic for discussion in the psychological literature, the phenomenal and psychological notions have not generally been carefully distinguished. Usually it is some aspect of awareness, such as introspection, attention, or self-consciousness, that psychological studies address. Even the psychological aspects of consciousness have had something of a bad name in psychology, at least until recently. Perhaps this is because of some unclarity in those notions, and the difficulties associated with high-level phenomena such as introspection. One might speculate that to a larger extent this bad name is due to their sharing a name with the notion of phenomenal consciousness, thus giving the appearance of partnership in crime.

One sometimes hears that psychological research has been "returning to consciousness" in recent years. What this seems to come to is that the psychological aspects of consciousness have been an active subject of research, with researchers becoming unafraid to use the term 'consciousness' for these. For the most part, however, phenomenal consciousness remains largely ignored. Perhaps this is understandable. While one can see how the methods of experimental psychology might lead to an understanding of the various kinds of awareness, it is difficult to see how these could provide an explanation of phenomenal experience.[13]

Cognitive models are well-suited to explaining psychological aspects of consciousness. There is no vast metaphysical problem in the idea that a physical system should be able to introspect its internal states, or that it should be able to rationally deal with information from its environment, or that it should be able to focus its attention first in one place and then in the next. It is clear enough that an appropriate functional account should be able to explain these abilities, although coming up with the correct account might take decades if not centuries. In general, however, these psychological properties are *all* that such psychological models explain, as we will see from an examination in Chapter 3.

[13]That the reversion to interest in consciousness does not include much room for phenomenal consciousness is indicated by examining the "topics of interest" listed for a new journal, *Consciousness and Cognition.* These include "EEG correlates of decision-making and awareness", "neuropathology of conscious experience and voluntary control", "priming studies of language processing", "selective and directed attention", "implicit memory", "blind sight", "the development of automaticity", "sub- and supraliminal signal detection", and "development of the self-concept". The only thing here that even suggests phenomenal consciousness is the second entry, and this is explicitly concerned only with "neuropathology". Conscious experience itself still seems to be off-limits for respectable psychological inquiry.

The central philosophical problem is that of phenomenal consciousness, and this is left untouched by the various explanations of psychological consciousness that have been put forward so far. In the following chapters, I will argue that there is a principled reason for this, and that the standard modes of explanation provided by cognitive science cannot succeed in explaining phenomenal consciousness.

In what follows, I will revert to using 'consciousness' to refer to phenomenal consciousness alone. When I wish to use the psychological notions, I will speak of 'psychological consciousness' or 'awareness'. It is phenomenal consciousness with which I will mostly be concerned.

Chapter 2

Supervenience and Explanation[1]

2.1 Supervenience

It is widely believed that the most fundamental facts about our universe are physical facts, and that all other facts are dependent on these. In a weak enough sense of 'dependent' this may be almost trivially true; in a strong sense, it is controversial. There is a complex variety of dependency relations between high-level facts and low-level facts in general, and the kind of dependency relation that holds in one domain, such as biology, may not hold in another, such as that of conscious experience. The philosophical notion of *supervenience* provides a unifying framework in which these dependence relations can be discussed. Using this notion, the explanatory and ontological status of consciousness with respect to the physical can be significantly clarified.

The notion of supervenience[2] formalizes the intuitive idea that one set of facts can fully determine another set of facts. The physical facts about the world seem to determine the biological facts, for instance, in that once all the physical facts about the world are fixed, there is no room for the biological facts to vary. This provides a rough characterization of the sense in which biological facts *supervene* on physical facts. In general, supervenience is a relation between two sets of properties:

[1]This chapter is not directly concerned with conscious experience, but rather builds a framework in which progress can be made in understanding the metaphysical and explanatory status of consciousness within the natural order. These preliminaries include an outline of various kinds of *supervenience*, which will be a fundamental notion in our discussion; an account of the notion of reductive explanation, whereby natural phenomena are explained in terms of more basic phenomena; a clarification of the notions of conceptual truth and logical possibility, which will play a significant role in the arguments; and an account of the relation between supervenience and reductive explanation. I also outline at length a picture of the metaphysical relationship between high-level phenomena and physical facts, one that seems to cover everything *except* consciousness.

Much of this material is philosophically technical. Some readers may wish to read only the material on supervenience in 2.1 for now, and perhaps glance at the characterization of reductive explanation in 2.2, before proceeding directly to the next chapter on consciousness, which should make sense independently of the rest of the material herein. The contents of this chapter can then be consulted to clarify some of the notions used in the later chapters, and to quell some worries that might have arisen.

B-properties—intuitively, the *high-level* properties—and A-properties, which are the more basic *low-level* properties.

For our purposes, the relevant A-properties will almost always be the physical properties: more precisely, the fundamental properties that make up the subject matter of a completed theory of physics. Perhaps these will include mass, charge, position; the properties of being an electron, a photon, and various other basic particles; properties characterizing the distribution of various spatiotemporal fields and the exertion of various forces; and so on. The precise nature of these properties is not important. If physics changes radically, the relevant class of properties may be quite different from those I mention, but the arguments will go through all the same. Such high-level properties as juiciness, lumpiness, giraffehood, and the like are excluded, even though there is a sense in which these properties are physical. In the future, talk of physical properties will be implicitly restricted to the class of fundamental properties unless otherwise indicated. I will sometimes speak of 'microphysical' or 'low-level physical' properties to be explicit.

The *A-facts* and *B-facts* about the world are the facts concerning the instantiation and distribution of A-properties and B-properties. So the physical facts about the world encompass all facts about the instantiation of physical properties at various locations in the spatiotemporal manifold. It will also be useful to stipulate that the world's physical facts include its basic physical laws. On some accounts, these laws are already determined by the totality of particular physical facts about the world, but we cannot take this for granted.

The template for our definition of supervenience is the following:

> B-properties supervene on A-properties if no two possible situations are indiscernible with respect to their A-facts while differing in their B-facts.

For instance, biological properties supervene on physical properties insofar as any two possible situations that are physically indiscernible are biologically indiscernible. More precise notions of supervenience can be obtained by filling in this template. Depending on whether one takes the "situations" in question to be individuals or entire worlds, one arrives at notions of *local* and *global* supervenience, respectively. And depending on how one construes the notion of possibility, one obtains notions of *logical* supervenience, *nomic* supervenience, and perhaps others.

[2]The notion of supervenience was introduced (not under that name) by Moore (1922). The name was introduced in print by Hare (1952). Davidson (1970) was the first to apply to the notion to the mind–body problem. More recently, a sophisticated theory of supervenience has been developed by Kim (1978; 1984), Horgan (1982; 1984c), Hellman and Thompson (1975), and others.

Local and global supervenience

B-properties *locally* supervene on A-properties if the A-properties of an individual determine the B-facts about that individual—that is, if any two possible individuals indiscernible with respect to their A-properties will be indiscernible with respect to all their B-properties. For example, shape locally supervenes on physical properties: any two objects with the same physical properties will necessarily have the same shape. On the other hand, value does not locally supervene on physical properties: an exact physical replica of the Mona Lisa is not worth as much as the Mona Lisa. In general, local supervenience of a property on the physical fails if that property is somehow *context-dependent*—that is, if an object's possession of that property depends not only on the object's physical constitution but also on its environment and its history. The Mona Lisa is more valuable than its replica because of a difference in their historical context: the Mona Lisa was painted by Leonardo, whereas the replica was not.[3]

B-properties *globally* supervene on A-properties, by contrast, if the A-facts about the entire *world* determine the B-facts: that is, if there are no two possible worlds indiscernible with respect to their A-properties, but whose B-properties differ. Local supervenience implies global supervenience, but not vice versa. For instance, it is plausible that biological properties globally supervene on physical properties, in that any world physically identical to ours would also be biologically indiscernible.[4] But they do not locally supervene. Two physically identical[5] organisms can differ in their biological characteristics. One might be fitter than the other, due to differences in their environmental contexts. It is even conceivable that physically identical organisms could belong to different species, if they had different evolutionary histories.

The distinction between global and local supervenience will not matter too much when it comes to consciousness, because it is very likely that insofar as consciousness supervenes on the physical at all, it supervenes locally. It seems unlikely that differences in environmental and historical contexts could prevent two physically identical creatures from having indistinguishable conscious experiences. Conscious experiences of creatures in our world seem to be dependent only on their internal structure. The role of the environment in fixing conscious experiences is entirely derivative on the role it plays in affecting internal structure. Witness the phenomena of hallucination and illusion, for instance.

---

[3]I assume, perhaps artificially, that individuals have precise spatiotemporal boundaries, so that their physical properties consist in the properties instantiated in that region of space-time.

[4]There is a slight complication, in that perhaps there could be a world just like ours physically, but with some extra non-physical ectoplasm that had biological properties of its own, so that that world differed biologically. I will discuss this complication below. For now, assume that we are only talking about worlds without extra "alien" properties that are nowhere exemplified in our world.

[5]I will use 'identical' in the sense of 'indiscernible' throughout, rather than in the sense of 'numerically identical'. Two crocodiles can be physically identical without being the same crocodile.

Logical and nomic supervenience

A more important distinction for our purposes is that between *logical* (or *conceptual*) supervenience, and mere *nomic* (or *empirical*) supervenience. B-properties *logically* supervene on A-properties if no two logically possible situations have the same A-properties but different B-properties.

I will say more about logical possibility in 2.4. For now, one can think of it loosely as possibility in the broadest sense, corresponding roughly to conceivability, quite unconstrained by the laws of our world. It is useful to think of a logically possible world that it would have been in God's power to create, had he so chosen.[6] God could not have created a world with male vixens, but he could have created a world with flying telephones. In determining whether it is logically possible that some statement is true, the only constraints are constraints of *meaning*. The notion of a male vixen is contradictory, so a male vixen is logically impossible; the notion of a flying telephone is conceptually coherent, if a little out of the ordinary, so a flying telephone is logically possible.

At the global level, biological properties logically supervene on physical properties. God could not have created a world that was physically identical to ours but biologically different. There is simply no logical space for the biological facts to independently vary. In a sense, when logical supervenience holds, *all there is* to the B-facts being as they are is that the A-facts be as they are. We might say that the A-facts are *constitutive* of the B-facts.

By contrast, B-properties *nomically* supervene on A-properties if the supervenience relation is a mere lawful correlation. We have nomic supervenience if, as a fact about our world, (1) clusters of A-properties are always accompanied by the same B-properties, and (2) this determination of B-properties by A-properties is not just a coincidence but lawful: that is, if there *were* to be any instance of such-and-such A-properties, it would be accompanied by the such-and-such B-properties. Such co-occurrence need not take place in every logically possible world: it need merely be a regular, reliable co-occurrence in our world.

An example: magnetic properties nomically supervene on electric properties in our world. Any two situations with the same electric properties will have the same magnetic properties; that is a consequence of the laws of our world. But this is arguably not a case of logical supervenience. It seems logically possible that there could be a world where objects had electric properties but no magnetic properties at all. The connection between them is merely a matter of the laws of our world. (The

[6]With one exception: God could not have created a world that was not created by God, even though a world not created by God is presumably logically possible. I will ignore this sort of complication.

question of whether electricity without magnetism is really a logical possibility is a subtle one; I bring the matter up only as an illustration.)

More precisely, B-properties nomically supervene on A-properties if any two *nomically possible* situations with indiscernible A-properties have the same B-properties. A nomically possible situation is one that is compatible with the laws of nature in our world. In general, nomic possibility and necessity is just possibility and necessity subject to the laws of nature,[7] and corresponds to what one may think of as real *empirical* possibility (as opposed to philosophers' mind-games, perhaps). A nomically possible situation is one that might come up in the real world, if the conditions were right. Such situations include past and future situations, along with all the counterfactually possible situations that might have come up in the world's history if boundary conditions had been different, or that might come up in the future, depending on how things go. So, B-properties nomically supervene on A-properties if *in fact*, given the way our world is, the A-properties of a situation determine its B-properties.

Logical supervenience clearly implies nomic supervenience. If A-properties determine B-properties in all logically possible situations, they certainly do so in all nomically possible situations. Nomic supervenience does not imply logical supervenience, however. The supervenience of magnetic properties on electric properties illustrates this.

It is hard to find cases of nomic supervenience *on the physical* without logical supervenience, for reasons that will become clear, but consciousness itself provides a promising illustration. Consciousness almost certainly nomically supervenes on physical properties (locally or globally), insofar as in the real world, any two physically indistinguishable creatures will have indistinguishable conscious experiences. However, it is not at all clear that consciousness logically supervenes on physical properties, as it seems *logically* possible that there could be a creature physically indistinguishable from me, but with no conscious experiences whatsoever, or with conscious experiences different from mine. (I will argue for this controversial conclusion in the next chapter, but for now it can serve as a convenient illustration.) If this is indeed the case, then the fact that identical creatures in practice have identical conscious experiences is simply part of the way the world is. Any necessity here is ensured only by the laws of nature, rather than by any logical or conceptual force.

The distinction between logical and nomic supervenience will be vital for our purposes.[8] It can be intuitively understood as follows. If B-properties logically supervene on A-properties, then once God (hypothetically) creates a world with certain

[7]A more informative name might be 'natural necessity', but this has an awkward adjectival form. 'Physical necessity' and 'causal necessity' are often used to pick out roughly the same brand of necessity, but I do not wish to beg the question of whether all the laws of nature are physical or causal. I will talk of 'nomic necessity' and sometimes 'empirical necessity' throughout.

A-facts, then the B-facts come along for free as an automatic consequence. If B-properties merely nomically supervene on A-properties, then after making sure of the A-facts, God will have to do more work in order to make sure of the B-facts: he will have to make sure there is a law relating the A-facts and the B-facts. Once the law is in place, the relevant A-facts will automatically bring along the B-facts; but one could, in principle, have had a situation where they did not.

One also sometimes hears talk of *metaphysical* supervenience, which is based on neither logical nor nomic necessity, but on "necessity *tout court*", or "metaphysical necessity" as it is sometimes known (drawing inspiration from Kripke's (1972) discussion of *a posteriori* necessity). I will argue later that metaphysical supervenience is just a variety of logical supervenience with an *a posteriori* semantic twist; for now, one may assume there is a notion of metaphysical supervenience, to be spelled out by analogy with the notions of logical and nomic supervenience above.[9]

The logical/nomic distinction and the global/local distinction cut across each other. It is reasonable to speak of both global logical supervenience and local logical supervenience, although we will more often be concerned with the former. When I speak of logical supervenience without a further modifier, the default assumption should be that global logical supervenience is being discussed. It is also coherent to speak of global and local nomic supervenience, but the nomic supervenience relations with which we are concerned will generally be local or at least localizable, for the simple reason that evidence for a nomic supervenience relation will generally consist in local regularities between clusters of properties.[10]

A problem with logical supervenience[11]

A problem with the notion of logical supervenience has already been alluded to and needs to be dealt with before we can proceed. This problem arises from the

[8]The distinction between logical and nomic supervenience is frequently glossed over or ignored in the literature, where the modality of supervenience relations is often left unspecified. Seager (1991) spells out a related distinction between what he calls *constitutive* and *correlative* supervenience. These correspond in a straightforward way to logical and nomic supervenience, although Seager does not analyze the notions in quite the same way.

[9]A notion of 'weak' supervenience is also sometimes mentioned, requiring only that "no B-difference without an A-difference" holds *within* a world, rather than across worlds (see Kim 1984 for details). This seems too weak to express an interesting dependency relation between facts. It is usually invoked either as a conceptual constraint on non-factual discourse (as in Hare 1952), in which case it is uninteresting for our purposes, or to express a kind of non-accidental within-world correlation that is not strictly necessary (as in Seager 1988), in which case nomic supervenience would serve much better.

[10]Global nomic supervenience without localized regularity is a coherent notion on a non-Humean account of laws, although perhaps not on a Humean (regularity) account. Even on a non-Humean account, though, it is hard to see what evidence for such a relation could consist in.

[11]This section is mostly of technical interest and can be skimmed or skipped.

logical possibility of a world physically identical to ours, but with additional non-physical stuff that is not present in our own world: angels, ectoplasm, and ghosts, for instance. Imagine a world with some extra angels hovering about, made of ectoplasm. These angels might conceivably have biological properties of their own, if they reproduced and evolved. Presumably the angels could have all sorts of beliefs, and their communities might have complex social structure.

The problem these examples pose is clear. If the angel world is logically possible, then there exist logically possible worlds that are physically indistinguishable but biologically different, so that biological properties cannot logically supervene on physical properties, by our definition. But we certainly *want* to say that biological properties are supervenient on physical properties. At the very least it seems that the biological properties of *this* world are determined by the physical properties, irrespective of the existence of other worlds where they are not.

This problem has caused some (e.g. Haugeland 1982 and Petrie 1987) to suggest that the relevant modality for the supervenience relations in question should be weaker than logical possibility and necessity, and that some weaker modality such as nomic possibility and necessity should be used instead. But this would be to give up the very useful distinction between logical and nomic supervenience outlined above, and also would be to ignore the fact that there is a very real sense in which the biological facts about our world are logically determined by the physical facts. Others (e.g. Teller 1984) have bitten the bullet by stipulating that worlds with extra non-physical stuff are not logically or metaphysically possible, despite appearances. This makes logical and metaphysical possibility seem quite arbitrary. However, it is possible to retain a useful notion of logical supervenience compatibly with the possibility of these worlds, as long as we fix the definition appropriately. (Horgan 1982 and Lewis 1983 address this problem in the context of discussing physicalism, but their solutions are somewhat different from mine.)

The solution to the problem is to turn supervenience into a thesis about *our* world (or, more generally, about particular worlds). This accords with the intuition that biological facts are logically determined by the physical facts in our world, despite the existence of bizarre worlds where they are not so determined. According to a revised definition, B-properties logically supervene on A-properties if the B-facts of our world are logically determined by the A-facts, in the following sense: in any possible world with the same A-facts, the same B-facts will hold. The existence of possible worlds with *extra* B-facts will thus not count against logical supervenience in our world, as long as *at least* the B-facts true in our world are true in all physically identical worlds. And this they generally will be (with an exception discussed below). If there is a koala eating in a gum tree in this world, there will be an identical koala eating in a gum tree in any physically identical world, whether or not that world has extra angels hanging around.

There is a minor complication. It might be argued that there are some biological facts about our world that do not hold in the angel world: the fact that our world has no living ectoplasm, for example, or the fact that all living things are based on DNA. We can get around this problem by ensuring that the relevant facts and properties are all particular facts and properties, and that universally quantified facts and properties are excluded.[12] From now on, it will be assumed that the relevant facts and properties do not include any with implicit or explicit universal quantifiers. For example, superlative properties such as that of being the most fecund organism in existence are excluded, as are the world-level universal properties given above.

Finally, it is desirable to make the logical supervenience of facts in our world a *lawful* thesis. If it were the case that there *would* have been non-physical living angels if things had gone a little differently in our world (perhaps a few different quantum fluctuations), even though the laws of nature were being obeyed, then it would be a mere accident of history that biological facts logically supervene on physical facts. One gets a stronger and more interesting metaphysical thesis by replacing the reference to "our world" in the definition of logical supervenience by a reference to "all nomically possible worlds", so that it is guaranteed that there will be no extra biological facts *however* the world evolves, as long as it obeys the laws of nature.

Our final definition of logical supervenience of B-facts upon A-facts therefore comes to this: for any nomically possible situation X, and any logically possible situation Y, if X and Y are indiscernible in their A-facts, then all the particular B-facts true of X are true of Y.[13] Or, more briefly: for any nomically possible situation, the particular B-facts about that situation are logically entailed by the A-facts about the situation. This definition works equally well for both local and global supervenience, where the "situations" in question are individuals and worlds respectively.

This definition captures the idea that supervenience claims are generally claims about our world, and nomically possible worlds in general, while retaining the key role of the logical modality. In what follows, these technicalities will be treated lightly. It will generally be clear that we are discussing nomically possible worlds, and the

[12]The crux here is that universally quantified facts and properties are *extrinsic to the world*, in a certain sense. They can be falsified merely by adding material to the world, while keeping everything else the same. They therefore do not supervene on any collection of particular facts about the world, and so certainly cannot supervene on a collection of particular physical facts. The relevant supervening facts and properties have to be restricted to those intrinsic to the world for supervenience to even get off the ground. There are arguably some non-universally-quantified facts that are also extrinsic in this sense—perhaps the very fact that our world is not angel world, or that it contains 29 (or whatever) different types of genetic structure. These can be handled similarly; in any case, the restriction to particular facts and properties excludes them.

[13]There is a parallel definition of 'metaphysical' supervenience. Of course, this problem does not arise for nomic supervenience, as ectoplasmic worlds are ruled out by the requirement of nomic possibility.

question at hand will be whether or not various particular facts are entailed by the physical facts about those worlds. Sometimes I will say something like "it is logically impossible that there could be a world physically identical to ours but different in its B-facts". Here, it should be understood that the kind of difference that is relevant is not whether or not the other world possesses extra particular facts, such as the facts that ectoplasm might bring along. Rather, a relevant difference would be a particular B-fact that holds in our world but not in the other world. 'Difference' should be understood as relevant difference in this sense. In addition, 'our world' will often be used as shorthand for 'a nomically possible world', at least when the context is a discussion of logical supervenience.

Supervenience and ontology

Logical and nomic supervenience have quite different ramifications for ontology (that is, for the matter of what there is in the world). If B-properties logically supervene on A-properties, then there is a sense in which once the A-facts are given, the B-facts are an ontological free lunch. Once God (hypothetically) made sure that all the physical facts in our world held, the biological facts came along for free. The B-facts merely redescribe what is described by the A-facts. They may be *different* facts (a fact about elephants is not a microphysical fact), but they are not *new* facts.

With mere nomic supervenience, the ontology is not so straightforward. To invoke the converse of a principle of Hume's, a contingent connection must connect distinct entities. In general, if B-properties are merely nomically supervenient on A-properties in our world, then there *could* have been a world in which our A-facts held without the B-facts. As we saw before, once, once God fixed all the A-facts, in order to fix the B-facts he had more work to do (this image is due to Kripke). The B-facts are something over and above the A-facts, and their satisfaction implies that there is something new in the world.

With this in mind we can formulate precisely the widely-held doctrine of *physicalism*, which is generally taken to hold that everything in the world is physical, or that there is nothing over and above the physical, or that the physical facts in a certain sense exhaust all the facts about the world. In our language, physicalism is true if all the facts about the world (globally) logically supervene on the physical facts. This captures the intuitive notion that if physicalism is true, then once God fixed the *physical* facts about the world, *all* the facts were fixed.

We have to use the notion of logical supervenience as developed in the previous section, of course, so that worlds with extra ectoplasmic facts do not count against physicalism in our world. To spell things out: physicalism is true if all the particular facts about our world are entailed by the physical facts, and if this is lawful. In more detail, physicalism is true if for every nomically possible world X and every logically possible world Y such that X and Y agree on their physical facts, then all the particular facts true of X are true of Y.[14] (Universally quantified facts are excluded for the reasons given above. The falsity of 'there are no angels' is compatible with the physical facts, but this does not affect the truth of physicalism.[15])

I will discuss this matter at much greater length in Chapter 4, where this definition of physicalism will be further justified. Some may object to the use of logical possibility rather than possibility *tout court*, or "metaphysical possibility". Those people may feel free to substitute metaphysical possibility for logical possibility in the definition above. We will see that it ultimately comes to the same thing.

2.2 Reductive explanation

The remarkable progress of science over the last few centuries has given us good reason to believe that that there is very little that is utterly mysterious about the world. For almost every natural phenomenon above the level of microscopic physics, there seems in principle to exist a *reductive explanation*, in the following sense: if we account for enough lower-level facts about the world, an explanation of the higher-level phenomenon will fall out.

Biological phenomena provide a clear illustration. Reproduction can be explained by giving an account of the genetic and cellular mechanisms that allow organisms to produce other organisms. Adaptation can be explained by giving an account of the mechanisms that that lead to appropriate changes in external function in response to environmental stimulation. Life itself is explained by explaining the various mechanisms that allow reproduction, adaptation, and the like. Once we have told the lower-level story in enough detail, any sense of fundamental mystery goes away: the phenomena that needed to be explained have been explained.

One can tell a similar story for most natural phenomena. In physics, we explain heat by telling an appropriate story about the energy and excitation of molecules. In astronomy, we explain the phases of the moon by going into the details of orbital motion and optical reflection. In geophysics, earthquakes are explained via an account of the interaction of subterranean masses. Even in cognitive science, it seems that to explain a phenomenon like learning, all we have to do is explain various functional mechanisms—the mechanisms that give rise to appropriate changes in behavior in response to environmental stimulation, at a first approximation. Many of the details

[14]This definition comes to much the same thing as definitions by Horgan 1982 and Lewis 1983, but unlike theirs it does not rely on the somewhat obscure notion of an "alien property", which Horgan and Lewis use to rule out ectoplasmic worlds from the range of relevant possible worlds.

[15]If we add to the supervenience base a second-order "that's all" fact, saying that all particular facts are entailed by the given set of particular physical facts, then even universally quantified facts will logically supervene. See 2.5 for more on this.

of these explanations currently evade our grasp, and are likely to prove highly non-trivial, but there is no *fundamental* mystery. We know that by finding out enough about the low-level story, the high-level story will fall out.

I will not precisely define the notion of reductive explanation until later, preferring to leave it vague and characterized by example. However, I can issue some brief caveats about what reductive explanation is not. A reductive explanation of a phenomenon need not require a *reduction* of that phenomenon, at least in some senses of the ambiguous notion of reduction. Multiply-realizable phenomena such as learning, for instance, are not reducible to lower-level phenomena in a certain sense of "reducible", but might nevertheless be reductively explainable in terms of lower-level phenomena. Reductive explanation is also not the be-all and end-all of explanation. There are many other sorts of explanation, some of which may shed more light on a phenomenon than a reductive explanation in a given instance. There are *historical* explanations, for instance, explaining the genesis of a phenomenon (such as life, say), where a reductive explanation only gives a synchronic account of how living systems function. There are also all sorts of *high-level* explanations, such as the explanation of some piece of behavior in terms of beliefs and desires. Even though this behavior might in principle be explainable reductively, a high-level explanation will often be much more comprehensible and enlightening. Reductive explanations should not be seen as displacing these other sorts of explanation. Each has its place.

Reductive explanation via functional analysis

What is it that allows such diverse phenomena as reproduction, learning, and heat to be reductively explained? In all these cases, the nature of the concepts involved is crucial. If someone objected to a cellular explanation of reproduction: "This explains how a cellular process can lead to the production of a complex physical entity that is similar to the original entity, but it doesn't explain *reproduction*", we would have little patience with them—for that is all that 'reproduction' *means*. In general, a reductive explanation of a phenomenon is accompanied by some rough-and-ready *analysis* of the phenomenon in question, whether implicit or explicit. The notion of reproduction can be roughly analyzed in terms of the ability of an organism to produce another organism in a certain sort of way. It follows that once we have explained the processes by which an organism produces another organism, we have explained that instance of reproduction.

This may seem to be a trivial point, but the possibility of this kind of analysis undergirds the possibility of reductive explanation in general. Without such an analysis, there would be no explanatory bridge from the lower-level physical facts to the phenomenon in question. With such an analysis in hand, all we need to do is to show how certain lower-level physical mechanisms might allow the analysis to be satisfied, and an explanation will result.

To be more specific, it seems that for the most interesting phenomena that require explanation, including phenomena such as reproduction and learning, the relevant notions can be analyzed *functionally*. Such notions can be characterized in terms of the performance of some function or functions (where 'function' is taken causally rather than teleologically), or in terms of the capacity to perform those functions.

It follows that once we have explained how those functions are performed, then we have explained the phenomena in question. Once we explain how an organism performs the function of producing another organism, we have explained reproduction, for all it means to reproduce is to perform that function. The same goes for an explanation of learning. All it means for an organism to learn, roughly, is for its behavioral capacities to adapt appropriately in response to environmental stimulation. If we explain how the organism is able to perform the relevant functions, then we have explained learning.

Explaining the performance of these functions is quite straightforward, in principle. As long as (1) the results of such functions are themselves characterizable physically, and (2) all physical events have physical causes, then there should be a physical explanation for the performance of any such function. One need only show how certain sorts of states are responsible for the production of appropriate resultant states, by a causal process in accord with the laws of nature. Of course the details of this kind of physical explanation can be non-trivial. Indeed, the details are what constitute the vast bulk of any reductive explanation, while the analysis component is often trivial. But in principle some story about low-level physical causation will explain how the relevant functions are performed, and therefore will explain the phenomenon in question.

Even a physical notion such as heat can be construed functionally: roughly, it is the kind of thing that expands metals, is caused by fire, leads to a particular sort of heat-perception, and the like. Once we have an account of how these various causal relations are fulfilled, then we have an account of heat. Heat is a *causal-role concept*, characterized in terms of what it is typically caused by and of what it typically causes, under appropriate conditions. Once empirical investigation tells us how the relevant causal role is played, the phenomenon is explained.

There are some technical complications here, but they are inessential. For example, 'heat' is not strictly *synonymous* with its characterization in terms of a causal role. In one sense, 'heat' can be seen to be equivalent to a characterization in terms of the motion of molecules, in that its reference across possible worlds is determined by this characterization rather than by the causal role. It remains the case, however, that *explaining* heat involves explaining the fulfillment of the causal role, rather than explaining the motion of molecules. To see this, note that the equivalence of 'heat' with 'the motion of molecules' is known *a posteriori*: we know this *as a result* of explaining heat. The concept of heat that we had *a priori*—before the phenomenon was explained—was roughly 'the thing that plays this causal role in the actual world'. Once we discover how that causal role is played, we have an explanation of the phenomenon. As a bonus, we know what heat *is*. It is the motion of molecules, as the motion of molecules is what plays the relevant causal role in the actual world. (Actually, other things also play this causal role, but we are simplifying.) There will be more on this technical point later.

A second minor complication is that many causal-role concepts are somewhat ambiguous between (a) the state that plays a certain causal role, and (b) the actual performance of that causal role. 'Heat' can be taken to denote either the molecules that do the causal work or the causal process (heating) itself. Similarly, 'perception' can be used to refer to either the act of perceiving, or the internal state that arises as a result. Nothing important for our purposes will turn on this. In general, an explanation of how the causal role is played will explain both (a) and (b).

A third complication is that many causal-role concepts are partly characterized in terms of their effect on our *phenomenology*: for instance, heat is naturally construed as the cause of our heat-sensations. Does this mean that we have to explain heat-sensations before we can explain heat? Of course, we do not have any good account of heat-sensations (or of phenomenology generally), so what happens in practice is that that part of the phenomenon is left unexplained. If we can explain how molecular motion comes about in certain conditions, and causes metals to expand, and stimulates our skin in certain ways, than the observation that this motion is *correlated* with our heat sensations is good enough. From the correlation, we infer that there is almost certainly a causal connection there. To be sure, no explanation of heat will be complete until we have an account of how that causal connection works, but the incomplete account is good enough for most purposes. It is somewhat paradoxical that we end up explaining almost everything about a *phenomenon* except for the details of how it affects our phenomenology, but it does not seem to be a problem in practice. It would not be a happy state of affairs if we had to put the rest of science on hold until we had a theory of consciousness.

Reductive explanations in cognitive science

The paradigm of reductive explanation via functional analysis works beautifully in cognitive science, as a rule. As we saw in the previous chapter, most non-phenomenal mental concepts can be analyzed functionally. Psychological states are characterizable in terms of the causal role they play. To explain these states, we explain how the relevant causation is performed.

In principle, one can do this by giving an account of the underlying neurophysiology. If we explain how certain neurophysiological states are responsible for the performance of the functions in question, then we have explained the psychological state. One need not always descend to the neurophysiological level, however.

One can frequently explain some aspect of mentality by exhibiting an appropriate *cognitive model*—that is, by exhibiting the details of the abstract causal organization of a system whose mechanisms are sufficient to perform the relevant functions, without specifying the physiochemical substrate in which this causal organization is implemented. In this way, one gives a *how-possibly* explanation of a given aspect of psychology, in that we have exhibited how the appropriate causal mechanisms *might* support the relevant mental processes. If we are interested in explaining the mental states of an *actual* organism or type of organism (e.g. learning in humans, as opposed to the possibility of learning in general), this sort of explanation must be supplemented with a demonstration that the causal organization of the model mirrors the causal organization of the organism in question.

Thus, to explain the possibility of learning, we can exhibit a model whose mechanisms lead to the appropriate changes in behavioral capacity in response to various kinds of environmental stimulation—a connectionist learning model, for example. To explain human learning, we must also demonstrate that such a model reflects the causal organization responsible for the performance of such functions in humans. The second step is usually difficult: we cannot exhibit such a correspondence directly, due to our ignorance of neurophysiology, so we usually have to look for indirect evidence, such as qualitative similarities in patterns of response, measurements of timing, and the like. This is one reason why cognitive science is currently in an undeveloped state. But as usual, the in-principle possibility of such explanation is a straightforward a consequence of the functional nature of psychological concepts.

Unfortunately, the kind of functional explanation that works so well for psychological states does not seem to work as an explanation of phenomenal states. The reason for this is straightforward. Whatever functional account of human cognition we give, there is a *further question*: Why is this kind of functioning accompanied by consciousness? No such further question arises for psychological states. If one asked about a given functional model of learning, "Why is this functioning accompanied by learning?", the appropriate answer is a trivial semantic answer: "Because all it *means* to learn is to function like this". There is no corresponding meaning analysis of consciousness that will do a similar job. Phenomenal states, unlike psychological states, are not defined by the causal roles that they play. It follows that explaining how some causal role is played is not sufficient to explain consciousness. The fact that consciousness accompanies the performance of a given function, if indeed it does, remains quite unexplained.

One can put the point this way. Given an appropriate functional account of learning, or of other psychological properties, it is simply *logically impossible* that something could instantiate that account without learning. However, no matter what functional account of cognition one gives, it seems entirely logically possible that that account could be instantiated without any accompanying consciousness. It might of course be empirically impossible—consciousness might in fact *arise* from that functional organization—but what is important here is that the notion is logically coherent.

If this is indeed logically possible, then any functional and indeed any physical account of mental phenomena will be fundamentally incomplete. To use a phrase due to Levine (1983), there is an *explanatory gap* between such accounts and consciousness itself. Even if the appropriate functional organization always gives rise to consciousness in practice, the question of *why* it gives rise to consciousness remains unanswered. I will explicate, justify, and develop this point later.

2.3 Logical supervenience and reductive explanation

The epistemology of reductive explanation meets the metaphysics of supervenience in a straightforward way. A phenomenon is reductively explainable in terms of some low-level properties precisely when it is logically supervenient on those properties. A phenomenon is reductively explainable in terms of physical properties—or simply "reductively explainable"—precisely when the phenomenon is logically supervenient on the physical. It is always global supervenience rather than local supervenience that is relevant here. To put things more carefully:

> A phenomenon is reductively explainable in terms of some lower-level properties iff the property of instantiating that phenomenon is globally logically supervenient on the low-level properties in question. A phenomenon is reductively explainable *simpliciter* iff the property of exemplifying that phenomenon is globally logically supervenient on physical properties.

This can be taken as something between an explication and a definition of reductive explanation. That our prior notion of reductive explanation implies (global) logical supervenience should be clear from the discussion above. If the property of exemplifying a phenomenon fails to logically supervene on some lower-level properties, then given any lower-level account of those properties, no matter how extensive, it will always be a *further question*: Why is this lower-level process accompanied by the phenomenon? Reductive explanation requires some kind of analysis of the phenomenon in question, where the low-level facts imply the realization of the analysis. So reductive explanation requires a logical supervenience relation. It is precisely because the phenomenon of reproduction is logically supervenient on lower-level facts, for instance, that it is reductively explainable in terms of those facts.

That logical supervenience *suffices* for reductive explainability is somewhat less clear. If a phenomenon P logically supervenes on some lower-level properties, then given an account of the lower-level facts L associated with any given instance of P, the exemplification of P is a logical consequence. Therefore an account of L will automatically yield an explanation of P, it seems. Why might this explanation nevertheless seem unsatisfactory? For two reasons. First, the lower-level facts L might be a vast hotch-potch of arbitrary-seeming details without any clear explanatory unity (an account of all the molecular motions underlying an instance of learning, for instance). Second, it is possible that different instances of P might be accompanied by very different sets of low-level facts, so that explanations of particular instances do not yield an explanation of the phenomenon as a type.

One option would be to hold that logical supervenience is merely *necessary* for reductive explanation, rather than sufficient. This is all that will be required for my arguments about consciousness in the next chapter. But it is more useful to note that there is *a* useful notion of reductive explanation such that logical supervenience is both necessary and sufficient. Instead of taking the problems above as indicating that the accounts in question are not *explanations*, we can instead take them to indicate that a reductive explanation is not necessarily a *satisfying* explanation. Reductive explanation, as I noted earlier, is not the be-all and end-all of explanation. Its role is chiefly to remove any deep sense of mystery surrounding some high-level phenomenon. It does this by reducing the bruteness and arbitrariness of the phenomenon in question to the bruteness and arbitrariness of some lower-level processes. The low-level processes in question may still be quite brute and arbitrary, as we have seen. Insofar as this is the case, a reductive explanation may not be very satisfying in giving us a deep understanding of a phenomenon, but it does eliminate any sense that there is something "extra" going on.

The gap between a reductive explanation and a satisfying explanation can generally be closed much further than this might suggest, however. This is due to two basic facts about the microphysics of our world: *autonomy* and *simplicity*. Microphysical causation and explanation seems to be autonomous, in that every physical event has a physical explanation; the laws of physics are sufficient to explain the events of physics on their own terms. Further, the laws in question are reasonably simple, so that the explanations in question have a certain compactness. Both of these things might have been otherwise. We might have lived in a world in which there were brutely emergent fundamental laws governing the behavior of high-level configurations such as organisms, giving rise to downward causation that overrode any relevant microphysical laws (the British emergentists, such as Alexander and Broad, believed in a world something like this; see McLaughlin 1992 for discussion). There might have been a world where in which the behavior of microphysical entities was governed only by a vast array of baroque laws, or perhaps a world in which microphysical behavior was lawless and chaotic. In worlds like these, there would be little hope of achieving a satisfying reductive explanation, as the bruteness of low-level accounts might never be simplified.

But our world, with its autonomy and simplicity on the lowest levels, seems to allow that sense can generally be made of even complex processes. The low-level facts underlying some high-level phenomenon often have some basic unity that allows for a comprehensible explanation. Given an instance of high-level causation, such as a trigger causing a gun to fire, we can not only isolate a bundle of lower-level facts that fix this causation; we can also tell a fairly simple story about how the causation is enabled, by encapsulating those facts under certain simple principles. This may not always work. It may be the case that some domains, such as those of sociology and economics, are far enough removed from the simplicity of low-level processes that their complexity prevents any satisfying reductive explanation from being carried out, even if they are logically supervenient. If so, then so be it: we can content ourselves with high-level explanations of those domains, while noting that logical supervenience implies that there is some kind of reductive explanation in principle, although perhaps only one that a superbeing could understand.

Note also that on this account reductive explanation is fundamentally *particular*, explaining particular instances of a phenomenon, without necessarily explaining all instances together. This is what one should expect. If some property is instantiable in vastly different sorts of ways, one would not expect a single explanation to cover all the instances. Temperature is instantiated quite differently in different media, for instance, and there are different sorts of explanation in each. At a higher level, it is most unlikely that there should be a single explanation covering all instances of murder. Still, there is frequently a certain unity across the explanation of particulars, in that a good explanation of one is often an explanation of many. This again seems to be a consequence of the underlying simplicity of our world, rather than a necessary property of explanation. In our world, the simple unifying stories that one can tell about lower-level processes often apply across the board, or at least across a wide range of particulars. It is also frequently the case, especially in the biological sciences, that the particulars have a common ancestry that leads to a similarity in the low-level processes involved. So the second problem mentioned above, that of unification of explanation across particular instances of a phenomenon, is not as much of a problem as it might be. Nevertheless, in this work I will mostly be concerned with the possibility of explanation of particulars.

There is much more that could be said about closing the gap between reductive explanation and satisfying explanation, but the matter deserves a lengthy tome in the philosophy of science in its own right, and it will not be too important for our purposes. What is most important is that (a) if logical supervenience fails (as I will argue it does for consciousness), then *any* kind of reductive explanation fails, even being generous about what counts as explanation. Also important is that (b) logical supervenience removes any residual *metaphysical* mystery about a high-level phenomenon, by reducing any brutality in that phenomenon to brutality in lower-level facts. Of secondary importance is that (c) if logical supervenience holds, then some sort of reductive explanation is possible. Although such explanations can fail to be satisfying or useful, this failure is not nearly as fundamental as the failure of explanation in domains where logical supervenience does not hold.

Further notes on reductive explanation

1. In general, any practical reductive explanation of a phenomenon will not go all the way to the microphysical level. That would be enormously difficult, giving rise to all the brutality problems that we have discussed. Instead, some high-level phenomena will be explained in terms of some properties at a slightly more basic level, as when reproduction is explained in terms of cellular mechanisms, or the phases of the moon are explained in terms of orbital motion. In turn, one hopes that the more basic phenomena will themselves be reductively explainable in terms of something more basic still. If all goes well, biological phenomena may be explainable in terms of cellular phenomena, which are explainable in terms of biochemical phenomena, which are explainable in terms of chemical phenomena, which are explainable in terms of physical phenomena. As for the physical phenomena, one tries to unify these as far as possible, but at some level physics has to be taken as brute: there may be no explanation of why the fundamental laws or boundary conditions are the way they are, for instance. This ladder of explanation is little more than a pipedream at the moment, but significant progress has been made. Given logical supervenience, along with the simplicity and autonomy of the lowest level, this sort of thing ought to be possible in principle. Whether the complexities of reality will make it practically infeasible is an open question.

2. According to the above account, it is at least conceivable that some phenomenon might be reductively explainable in terms of some lower-level properties without being reductively explainable *simpliciter*. This might happen in a situation where C-properties logically supervene on B-properties, and are therefore explainable in terms of B-properties, but where B-properties themselves are not logically supervenient on the physical. There is clearly a sense in which such an explanation is reductive, and a sense in which it is not. For the most part, I will be concerned with reductive explanation in terms of the physical, or in terms of properties that are themselves explainable in terms of the physical, and so on. Even if the C-properties above are reductively explainable in a relative sense, their very existence presupposes the failure of reductive explanation in general.

3. Note that *local* logical supervenience is too stringent a requirement for reductive explanation. One can reductively explain the exemplification of even context-dependent properties of an individual by giving an account of how relevant environmental relations come to be satisfied, as well as of the internal structure of the individual. As long as some phenomenon is globally supervenient, there will be some range of lower-level facts, perhaps wide in their spatiotemporal extent, that will in principle yield an explanation of the phenomenon in question.

4. In principle, there will be two projects in reductive explanation of a phenomenon like life, learning, or heat. First, there is a project of *explication*, where we clarify just what it is that needs to be explained, by means of some analysis: learning might be analyzed as a certain kind of adaptational process, for instance. Second, there is a project of *explanation*, where we see how that analysis comes to be satisfied by the low-level facts. The first project is conceptual, and the process is empirical. For many or most phenomena, the conceptual stage will be quite trivial. For some phenomena, however, such as belief, explication can be a major hurdle in itself. In practice, of course, there is never a clean separation between the projects, as explication and explanation take place in parallel.

I will argue shortly that most phenomena in our world are indeed logically supervenient on the physical, and are therefore reductively explainable in principle. We can see how this might work for various kinds of phenomena, such as the functionally characterizable phenomena discussed earlier. To explain an instance of a functional property, we account for low-level facts sufficient to entail that there will be an entity playing the relevant functional role, given the laws of physical causation. To explain a toaster, for instance, we give a physical account of the process by which it causes bread to be toasted, when placed inside a toaster. Such a lower-level story is straightforward in principle. By telling a physical story that entails the toasting of bread in this way, we reductively explain a toaster. So it goes elsewhere, with the reductive explanation of life, learning, and any other functionally characterizable phenomenon. Wherever we have logical supervenience, we have the possibility of reductive explanation.

2.4 Conceptual truth and necessary truth[16]

In the above I have relied heavily on the notions of logical possibility and necessity. It is now time to say something more about this. One way of understanding the logical

[16]This section deals with technical issues in the philosophy of language. It can be skipped by non-philosophers without too much loss, although a glance at the beginning of the third section on *a posteriori* necessity would be useful in opening doors to some later material.

necessity of a statement is as truth in virtue of meaning: a statement is logically necessary if its truth is ensured by the meaning of the concepts involved. It is also possible to explicate the logical necessity of a statement in terms of its truth across possible worlds, although that requires some care; I will discuss this later in this section. The notion of logical necessity is not to be identified with some narrow notion involving derivability in first-order logic, or some other syntactic formalism; I take it that justification of the axioms and rules of these formalisms precisely depends on their logical necessity in the broader sense.

This talk clearly requires taking seriously the notion of *conceptual truth*—that is, the notion that some statements are true or false simply by virtue of the meaning of the terms involved. The technical apparatus of logical supervenience depends on this, and so do the examples I have given of reductive explanation. All these relied on characterizations of the concepts involved. A reductive explanation of reproduction, for instance, was justified by arguing that low-level details entailed that certain functions would be performed, and that performance of these functions is all that reproduction *means*.

The notion of conceptual truth has had a bad name in some circles since the critique by Quine (1951), who argued that there is no useful distinction between conceptual truths and empirical truths. The objections to these notions usually cluster around the following points: (1) Most concepts do not have definitions giving necessary and sufficient conditions (this observation has been made many times but is most frequently attributed to Wittgenstein 1953); (2) Most apparent conceptual truths are in fact revisable, and could turn out to be false in the face of sufficient empirical evidence (this point is due to Quine); and (3) Considerations about *a posteriori* necessity, due to Kripke (1972), show that application-conditions of many terms across possible worlds cannot be known *a priori*.

These considerations count against an overly simplistic view of conceptual truth, but they do not affect the way in which I am using these notions. In particular, it turns out that the class of *supervenience conditionals* – 'if the A-facts about a situation are X, then the B-facts are Y', where the A-facts fully specify a situation—are unaffected by these considerations. These will be the only conceptual truths that my arguments will ultimately need, and none of the considerations above count against them, as we see shortly. Later in this section I will also analyze the relationship between conceptual truth and necessary truth in more detail, and spell out the role these play in understanding logical supervenience.

1. Definitions

The absence of cut-and-dried definitions is the least serious of the three difficulties with conceptual truth. None of my arguments depend on the existence of such

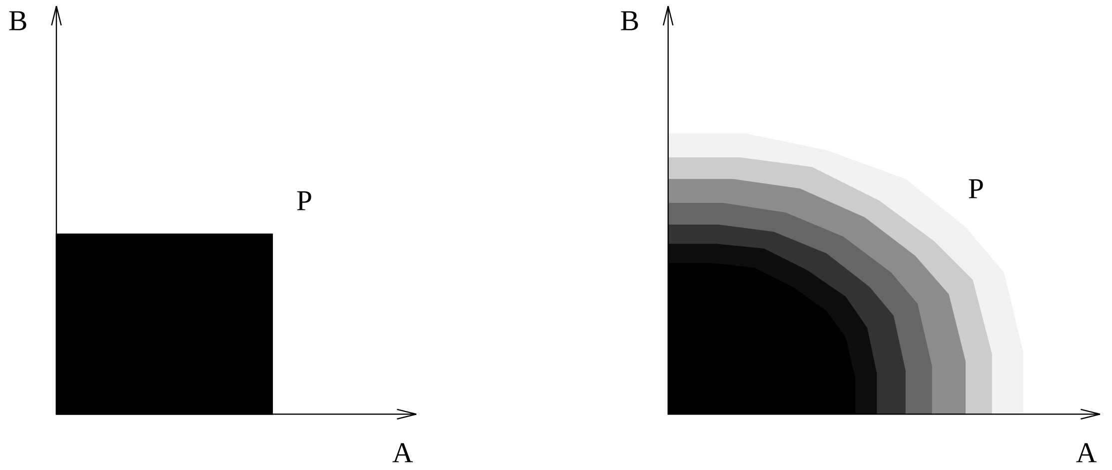

Figure 1: Two ways in which a property P might depend on properties A and B.

definitions. I occasionally rely on analyses of various notions, but these analyses need only be rough-and-ready, without any pretense at providing precise necessary and sufficient conditions. Most concepts (e.g., 'life') are somewhat vague in their application, and there is little point trying to remove that vagueness via arbitrary precision. Instead of saying "a system is alive if and only if it reproduces (with weight 0.65) and adapts with utility 800 or greater (with weight 0.85), and/or metabolizes with efficiency 75%", we can simply note that if a system exhibits these phenomena to a sufficient degree then it will be alive, by virtue of the meaning of the term. If an account of relevant low-level facts fixes the facts about a system's reproduction, utility, metabolism, and so on, then it also fixes the facts about whether the system is *alive*, insofar as that matter is factual at all.

We can sum this up with a schematic diagram (Figure 1) showing how a high-level property P might depend on two low-level parameters A and B, each of which can take on a range of values. If we had a crisp definition in terms of necessary and sufficient conditions, then we would have something like the picture at left, where the dark rectangle represents the region in which property P is instantiated. Instead, the dependence is invariably something like the picture at right, where the boundaries are vague and there is a large area in which the matter of P-hood is indeterminate, but there is also an area in which the matter is clear. (It may be indeterminate whether bacteria or computer viruses are alive, but there is no doubt that dogs are alive.) Given an example in the determinate area, exemplifying A and B to sufficient degrees that P is exemplified, the conditional 'if x is A and B to this degree, then x is P' is a conceptual truth, despite the lack of a clean definition of P. Any indeterminacy in such conditionals, in the gray areas, will reflect indeterminacy in the facts of the matter, which is as it should be. The picture can straightforwardly be extended to

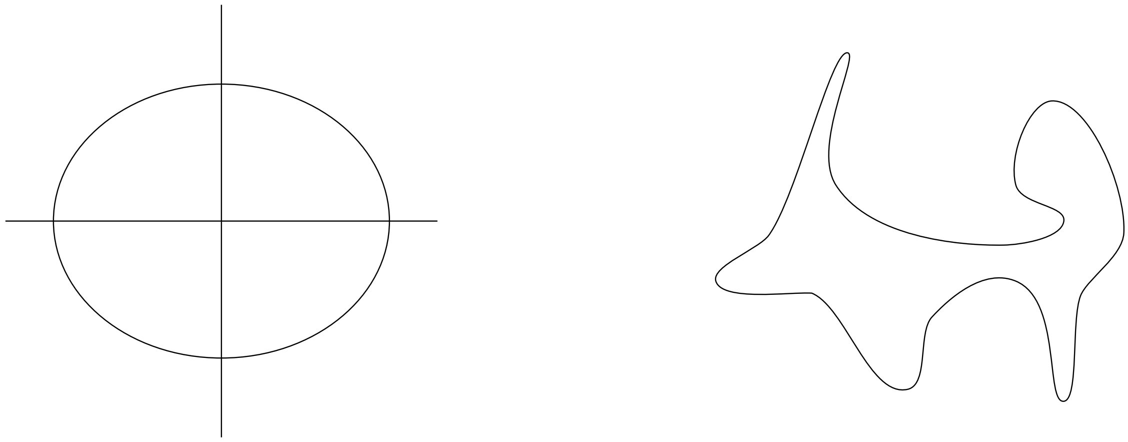

Figure 2: The round curve $2x^2 + 3y^2 = 1$, and non-round friend.

dependence of a property on an arbitrary number of factors, and to supervenience conditionals in general.

Importantly, then, certain A-facts can *entail* certain B-facts without there being a clean definition of B-notions in terms of A-notions. The above case provides an example: there is no simple definition of P in terms of A and B, but the facts about A and B in an instance entail the facts about P. For another example, think about the *roundness* of closed curves in two-dimensional space (Figure 2). There is certainly no perfect definition of roundness in terms of simpler mathematical notions. Nevertheless, take the figure at left, specified by the equation $2x^2 + 3y^2 = 1$. There is a fact of the matter—this figure is round— insofar as there are ever facts about roundness at all (compare to the figure at right, which is certainly not round). Further, this fact is *entailed* by the basic description of the figure in mathematical terms— given that description, and the concept of roundness, the fact that the figure is round is determined. Given that A-facts can entail B-facts without a definition of B-facts in terms of A-facts, the notion of logical supervenience is unaffected by the absence of definitions.

2. Revisability

The second objection, due to Quine (1951), is that purported conceptual truths are always subject to revision in the face of sufficient empirical evidence. For instance, if evidence forces us to revise various background statements in a theory, it is possible that a statement that once appeared to be conceptually true might turn out to be false.

This is so for many purported conceptual truths, but it does not apply to the supervenience conditional that we are considering, which have the form 'if the low-level facts turn out like this, then the high-level facts will be like that'. The facts specified in the antecedent of this conditional effectively include all relevant empirical factors. Empirical evidence could show us that the antecedent of the conditional is false, but not that the conditional is false. In the extreme case, we can ensure that the antecedent gives a full specification of the low-level facts about the world. The very comprehensive of the antecedent ensures that empirical evidence is utterly irrelevant to the conditional's truth-value. (This picture is complicated by the fact that empirical developments about the actual world can affect judgments about counterfactual possible worlds. I will deal with such matters shortly. For now, we are only concerned with conditionals about the way the actual world turns out.)

While Quine's critique provides a plausible argument that there are not many *short* conceptual truths, nothing in the critique counts against the constrained, complex sort of conceptual truth that I have been considering. The upshot of Quine's argument is that the truth-conditions of a high-level statement may not be easily *localizable*, as all sorts of factors might have some kind of indirect relevance, but it does not count against the global truth-conditions provided by a supervenience conditional. Indeed, if meaning determines a function from possible worlds to reference classes, and if possible worlds are finitely describable (say, in terms of arrangement of basic qualities in those worlds), then there will automatically be a vast class of conceptually true conditionals that result.

3. *A posteriori* necessity

It has traditionally been thought that all conceptual truths are knowable *a priori*, as are all necessary truths, and that the classes of *a priori* truth, necessary truth, and conceptual truth are closely related or even coextensive. Kripke's *Naming and Necessity* (1972) threw a wrench into this picture by demonstrating the existence of a large class of necessarily true statements whose truth is not knowable *a priori*. An example is the statement 'water is H_2O'. We certainly cannot know this to be true *a priori*; for all we know (or for all we knew at the beginning of inquiry), water is made of XYZ. Kripke argues compellingly that nevertheless, given that water is H_2O in the actual world, then water is H_2O in all possible worlds. It follows that 'water is H_2O' is a necessary truth despite its *a posteriori* nature.

This raises a few difficulties for us. For instance, on some accounts, these necessary truths are conceptual truths, implying that not all conceptual truths are knowable *a priori*. On alternative accounts, such statements are not concepual truths, but then the link between conceptual truth and necessity is broken. I will spend some time setting up a systematic framework for dealing with these issues, which will recur. It turns out that these complications do not change anything fundamental to our arguments, but it is worth taking the trouble to get very clear about what is going on.[17]

On the traditional view of reference, derived from Frege although cloaked here in modern terminology, a concept determines *a priori* a function $f : W \rightarrow R$ from possible worlds to referents. Call such a function an *intension*; together with a specification of a world w, it determines an *extension* $f(w)$. The Fregean picture associated a single intension with any given concept. This intension, the *sense* of the concept, was supposed to determine the reference of the concept, and could be thought of as the *meaning* of the concept in question.

More recent work has recognized that that no single intension can do all the work that a meaning needs to do. The picture developed by Kripke (1972) complicates things by noting that reference in the actual world and in counterfactual possible worlds is determined by quite different mechanisms. Something like a Fregean sense survives in this picture as the *prior intension* whereby reference is fixed in the *actual* world, depending on how the world turns out. As in (2) above, this intension must be determined *a priori*, as it specifies how we will react to empirical developments, and so cannot depend on those developments. For instance, the actual-world reference of 'water' might be fixed by an intension along the lines of 'the clear, drinkable liquid found in our environment', or 'watery stuff' for short.[18]

[17]The framework outlined in what follows is a synthesis of ideas due to Kripke, Kaplan, Stalnaker, Lewis, Evans, Davies and Humberstone, and others who have addressed the two-dimensional nature of reference and necessity. All sorts of interesting technical issues and complications arise when spelling out the picture I am advocating in more detail. I hope to address those issues in the philosophy of language in much greater detail elsewhere. Here, I am spelling out just enough of the picture to handle the explanatory and metaphysical issues with which we will be concerned; it turns out that the details I am passing over are largely irrelevant to those concerns. I also believe that the two-dimensional picture has great power in addressing issues as diverse as meaning holism, sense and reference, and the role of belief contents in psychological explanation. These applications require an extended treatment in their own right, however.

[18]By specifying a prior intension with an English sentence, I am not suggesting that intensions are linguistically represented. I am merely concerned to flesh out the character of the relevant function from possible worlds to extension.

A brief characterization of a prior intension as something like "the clear drinkable liquid found in our environment" is inevitably a simplification. To get a better idea of the intension, we can reflect on what we would say or would have said in reaction to various empirical developments concerning water. For instance: if it had turned out that the liquid in lakes was H_2O and the liquid in oceans XYZ, then we would probably have said both were water; if the stuff in oceans and lakes was a mixture of 95% A and 5% B, we would probably have said that A but not B was water; if a substance neither clear or drinkable was found to bear an appropriate microphysical relation to the clear, drinkable liquid, then we would call that substance 'water' too, and so on. The full conditions for actual-world waterhood will be quite vague at the edges, may not be immediately apparent on reflection, and are best regarded as being *a priori* to the community as a whole rather than to any individual. None of this makes an essential difference to the picture I am describing, however; what

The prior intension does not suffice to determine reference in counterfactual possible worlds, however. 'Water', for instance, picks out H_2O in all possible worlds, given that it picks out H_2O in the actual world. In a world where the dominant clear, drinkable liquid is XYZ rather than H_2O, this liquid is not water; it is merely watery stuff. Counterfactual reference is determined by a *posterior intension*, which is not knowable *a priori* as it depends on how *our* world turns out. However, it has a close relation to the prior intension above. It is determined by firstly evaluating the prior intension at the actual world, and secondly *rigidifying* this evaluation so that the same class is picked out in all possible worlds. Given that the prior intension 'watery stuff' picks out H_2O in the actual world, it follows from rigidification that H_2O is the referent in all possible worlds.

There is sometimes a tendency to suppose that *a posteriori* necessity of this kind makes *a priori* conceptual analysis irrelevant. But this is clearly false. Before we even get to the point of rigidification, there is a conceptual story to tell about what makes an actual-world X *qualify* as an X in the first place. This story can only be told by an analysis of the prior intension.

We can sum this up by saying 'water' is conceptually equivalent to '*dthat*(watery stuff)', where *dthat* is a version of Kaplan's rigidifying operator, converting an intension into a rigid designator by evaluation at the actual world (Kaplan 1979). The single Fregean intension has been fragmented into two: a prior intension ('watery stuff') that fixes reference in the actual world, and a posterior intension ('H_2O') that picks out reference in counterfactual possible worlds, and which depends on how the actual world turned out. The prior intension corresponds to what Kaplan (1978) has

---

is important is that this intension cannot be dependent on *a posteriori* developments, as it specifies our reaction to those developments.

The conditions involved in the prior intension of 'water' may centrally involve a causal connection between the kind in question and the community's use of the term 'water', and may therefore be quite unlike a straightforward description of water's manifest properties. The picture I am outlining is therefore neutral between 'causal' and 'descriptive' theories of reference (although surely the correct picture of reference-fixation will combine insights from both theories, perhaps along the lines of the "causal descriptivism" advocated by Lewis 1984.)

The most intricate problems with prior intensions involve terms used by individuals who do not have a full grasp of the concept involved, but who nevertheless succeed in reference by deference to the grasp of others in the community; e.g. my term 'elm' picks out something different from my term 'beech', even though I know nothing about the difference between elms and beeches aside from the difference in their names. These matters can be handled in the current framework (an individual's prior intensions are often grounded in those of an entire community, where the name provides a link), but in any case they are irrelevant to the explanatory and metaphysical issues with which we will be concerned. For our purposes, we may as well assume that every individual in a community is as well-informed as everyone else about what it takes for something to qualify as water, life, or whatever; alternatively, we can pretend that the community is a giant individual.

called the *character* of a term;[19] the posterior intension corresponds to its *content.*

Both the prior and posterior intensions can be seen as functions $f : W \rightarrow R$ from possible worlds to reference classes, where the possible worlds in question are seen in subtly different ways. The prior intension takes us from a possible world seen as a *context of utterance* (in Kaplan's term) to a reference class picked out in that world. The posterior intension takes us from a possible world seen as a *circumstance of evaluation* to a reference class in that world, where the circumstance of evaluation may be quite different from the content of utterance.

There is a slight asymmetry in that the context of utterance but not the circumstance of evaluation is what Quine (1969) calls a *centered* possible world. This is an ordered pair consisting of a world and an *index* representing the spatiotemporal position within that world of an agent using the term in question (Quine suggests that the index be a point in space-time; Lewis (1979) suggests that it might more appropriately be an individual within the world, or a pair of an individual and a time). Such an index is necessary to capture the fact that a term like 'water' picks out a different reference class for me than for my twin on Twin Earth, despite the fact that we live in the same universe[20] . It is only our position in the universe that differs, and it is this position that makes a relevant difference to the reference-fixing process.

This phenomenon arises in an especially obvious way for indexical terms such as 'I' (see Kaplan 1989; Lewis 1979; Perry 1979), whose reference is clearly dependent on the agent using the term and not just on the overall state of the world. There is a less overt indexical element in notions such as 'water', however, which can be roughly analyzed as 'the dominant clear, drinkable liquid *in our environment*'. It is this indexical element that requires prior intensions to depend on centered possible worlds. Once actual-world reference is fixed, however, no special index is needed to evaluate reference in a counterfactual possible world. The circumstance of evaluation can therefore be represented by a simple possible world without an index.

[19]Kaplan's notions of character and content are developed to deal with indexical and demonstrative terms such as 'I' and 'that', but they can be seen to apply more generally. Kaplan presents character as a function from context to content, or in our terms, as a function from a centered possible world to a posterior intension. The posterior intension in question is generally a projection of that individual across possible worlds, however, so this comes to much the same thing as a function from a centered possible world to an individual. I will use functions of the latter sort, for purposes of symmetry and simplicity.

[20]Arguably, the prior intension is only well-defined over possible worlds centered on individuals thinking an appropriate thought, or making an appropriate utterance; in the extreme, one might argue that the prior intension of my thought is only determinate across worlds centered on a psychological duplicate of me. The prior intension is naturally extendible to a wider class of worlds, however. In any case, it does not matter if the intension yields indeterminate reference at many worlds.

All this can be formalized by noting that reference in a possible world is not determined solely by a singly-indexed function $f : W \rightarrow R$. Instead, reference in a possible world depends both on that possible world and on the way the actual world turns out. That is, a concept determines a doubly-indexed function

$$F : W^* \times W \rightarrow R,$$

where W^* is the space of centered possible worlds and W is the space of ordinary possible worlds. The first parameter represents contexts of utterance, or ways the actual world might turn out, whereas the second parameter represents circumstances of evaluation, or counterfactual possible worlds. Equivalently, a concept determines a family of functions

$$F_a : W \rightarrow R$$

for each $a \in W^*$ representing a way the actual world might turn out, where $F_a(w) = F(a, w)$. If a is a world in which watery stuff is H_2O, then F_a picks out H_2O in any possible world. Given that in our actual world water *did* turn out to be H_2O, this F_a specifies the correct application-conditions for 'water' across counterfactual worlds. If our world had turned out to be a different world b in which watery stuff was XYZ, then the relevant application-conditions would have been specified by F_b, a different intension which picks out XYZ in any possible world.

The function F is determined *a priori*, as all *a posteriori* factors are included in its parameters. From F we can recover both of our singly-indexed intensions. The prior intension is the function $f : W^* \rightarrow R$ determined by the "diagonal" mapping $f : w \mapsto F(w, w)$, the function whereby reference in the actual world is fixed. (Strictly speaking, this should be $F(w, w')$ where w' is the possible world derived by removing the "center" from the centered possible world w. I will ignore this sort of nicety.) The posterior intension is the mapping $F_a : w \mapsto F(a, w)$, where a is fixed to be our actual world. This intension picks out reference in counterfactual worlds.[21]

In the reverse direction, the doubly-indexed function F and therefore the posterior intension F_a can be derived from the prior intension f, with the help of a projection operator $D : R \times W \rightarrow R$, which goes from a class picked out in some world to members of "that" class in another possible world. The relevant relationship is $F(a, w) = D(f(a), w)$. The posterior intension F_a is just the function $D(f(a), \_)$, which we can think of as *dthat* applied to the intension given by f. The fact that $f(w) = F_w(w)$—that is, that the prior and posterior intensions coincide in their application to the actual world—falls out as an immediate consequence, as $D(f(w), w)$ is straightforwardly $f(w)$.[22]

[21]See Stalnaker 1978 for an outline of a "two-dimensional" framework much like this one.

[22]I have skimmed over some details. References classes R are perhaps best viewed as subsets of

For some expressions, derivation of the posterior intension from the prior intension will be even easier. These are expressions such as 'watery stuff' that refer by "description" rather than by "rigid designation", with no implicit 'dthat' in the corresponding concept. These expressions apply to counterfactual worlds independently of how the actual world turns out. We therefore need no projection operator, and the posterior intension is more or less the same as the prior intension (except for differences due to centering in the worlds in question). The framework I have outlined can handle both sorts of concepts.

The prior and posterior intensions f and F_a correspond to Kaplan's character and content respectively. Both of these have been put forward as candidates for the "meaning" of a concept: Kaplan reserves the honorific for the former, whereas Putnam (1975) leans toward the latter. It seems to me that there is not really a substantial issue here: we might as well call the character and content the *a priori* and *a posteriori* aspects of meaning respectively.

It follows that both of these intensions back a certain kind of conceptual truth, or truth in virtue of meaning. The prior intension backs *a priori* conceptual truths, such as 'water is watery stuff'. Such a statement will be true no matter how the actual world turns out, although it need not hold in all non-actual possible worlds. The posterior intension does not back *a priori* truths, but backs truths that hold in all counterfactual possible worlds, such as 'water is H_2O'. Both varieties qualify as truths in virtue of meaning; they are simply true in virtue of different aspects of meaning.

It is also possible to see both as varieties of *necessary* truth. The latter corresponds to the more standard construal of a necessary truth. The former, however, can also be construed as truth across possible worlds, as long as these possible worlds are construed as contexts of utterance, or as *ways the actual world might turn out*. On this subtly different construal, a statement S is necessarily true if no matter how the actual world had turned out, we would have said that S was true (or: an utterance of S, made in that world, would have been true). If the actual world had turned out to be a world in which watery stuff was XYZ, then 'XYZ is water' (uttered in the actual world) would have been true. So, according to this construal on which possible worlds are *considered as actual*, 'water is watery stuff' is a necessary truth.

This kind of necessity is what Evans (1979) calls "deep necessity", as opposed to "superficial" necessities like 'water is H_2O'. It is analyzed in detail by Davies and

---

the set of entities in a world w, so that mentions of R should be replaced by R_w. Also, I have avoided explicitly noting the dependence of the functions f and F on the relevant concept C, as another subscript would be gruesome. It should be noted that it is possible that the function D might also depend on C, although it is a subtle question whether entities are projected differently across possible worlds depending on the sortal concept used to pick out their reference class.

Humberstone (1980) by means of a modal operator they call "fixedly actually". Deep necessity, unlike superficial necessity, is unaffected by *a posteriori* considerations. These two varieties of possibility and necessity apply always to *statements*. There is only one kind of possibility of *worlds*; the two approaches differ on how the truth of a statement is evaluated in a world.

We can see this in a different way by noting that there are two sets of *truth-conditions* associated with any statement. If we evaluate the terms in a statement according to their prior intensions, we arrive at the *prior* truth-conditions of the statement; that is, a set of centered possible worlds at which the statement, evaluated according to the prior intensions of the terms therein, turns out to be true. The prior truth-conditions tell us how the actual world has to be for an utterance of the statement to be true in that world; that is, they specify those *contexts* in which an utterance of the statement will be true. For instance, the prior truth-conditions of "water is wet" specify roughly that such an utterance will be true in the set of worlds in which watery stuff is wet.

If instead we evaluate the terms involved according to their posterior intensions, we arrive at the more familiar *posterior truth-conditions*. These conditions specify the truth-value of a statement in counterfactual worlds, given that the actual world turned out as it did. For instance, the posterior intension of 'water is wet' (uttered in this world) specify roughly those worlds in which H_2O is wet. Note that there is no danger of an ambiguity in actual-world truth: the prior and posterior truth-conditions will always specify the same truth-value when evaluated at the actual world.

If we see a proposition as a function from possible worlds to truth-values, then these two sets of truth-conditions yield two *propositions* associated with any statement. The prior intensions of the terms involved lead to a *prior proposition*, which holds in precisely those contexts of utterance in which the statement expresses a truth. (This is the "diagonal proposition" of Stalnaker 1978. Strictly speaking it is a centered proposition, or a function from centered possible worlds to truth-values.) The posterior intensions yield a *posterior* proposition, which holds in those counterfactual circumstances in which the posterior truth-conditions are satisfied. The posterior proposition is Kaplan's "content" of an utterance and is more commonly seen as the proposition expressed by a statement, but the prior proposition is also central.[23]

[23]It is arguable that prior propositions are more natural candidates for the object of a *belief* than posterior propositions. Certainly the prior proposition seems more naturally to reflect the cognitive role played by a belief. The prior proposition associated with a belief will be determined entirely by the internal state of the believer, and so is so a natural candidate for the "narrow content" of the belief, which many have suggested is more central to psychological explanation than the "wide content", corresponding to the posterior proposition and dependent on the state of the environment. See Fodor 1987, White 1982, and Loar 1988 for related suggestions, although see Stalnaker 1990 for some worries.

The two kinds of necessary truth of a statement correspond precisely to the necessity of the two kinds of associated proposition. A statement is necessarily true in the *a priori* sense if the associated prior proposition holds in all centered possible worlds (that is, if the statement expresses a truth in any context of utterance). A statement is necessarily true in the *a posteriori* sense if the associated posterior proposition holds in all possible worlds (that is, if the statement as uttered in the *actual* world is true in all counterfactual worlds). The first corresponds to Evans' deep necessity, and the second to the more familiar superficial necessity.

To illustrate, take the statement 'water is H_2O'. The prior intensions of 'water' and 'H_2O' differ, so that we cannot know *a priori* that water is H_2O; the associated *prior* proposition is not necessary. Nevertheless, the posterior intensions coincide, so that 'water is H_2O' is true in all possible worlds when evaluated according to the posterior intensions— that is, the associated *posterior* proposition is necessary. Kripkean "*a posteriori* necessity" arises just when the posterior intensions in a statement back a necessary proposition, but when prior intensions do not.

Consider by contrast the statement 'water is watery stuff'. Here the associated prior intensions $f : W^* \to R$ of 'water' and 'watery stuff' are the same, so that we can know this statement to be true *a priori*, as long as we possess the concept of water. The associated prior proposition is necessary, so that this statement is necessarily true in Evans' "deep" sense. However, the posterior intensions $F_a : W \to R$ differ (due to an implicit "dthat" in the analysis of 'water' but not 'watery stuff')—in a world where XYZ is the clear, drinkable liquid, the posterior intension of 'watery stuff' picks out XYZ but that of 'water' does not. The associated *posterior* proposition is therefore not necessary, and the statement is not a necessary truth in the more familiar sense; it is an example of Kripke's "contingent *a priori*".

Note that this account justifies a link between conceptual truth and necessary truth, one that is sometimes held to be questionable on the basis of Kripkean considerations. There are two varieties of conceptual truth, depending on whether the "concepts" correspond to prior or posterior intensions; and there are two varieties of necessary truth, depending on a similar decision. As long as one makes parallel decisions in the two cases, a statement is conceptually true if and only if it is necessarily true. There are parallel senses of 'conceptual truth' and 'necessary truth' on which 'water is H_2O' is both conceptually and necessarily true; and there are distinct parallel senses in which 'water is watery stuff' is both conceptually and necessarily true. I will discuss this link in more detail shortly.

In general, apparent problems that arise from these Kripkean considerations are usually a consequence of trying to squeeze the doubly-indexed picture of reference into a single notion of meaning or of necessity. Such problems can usually be dissolved by explicitly noting the two-dimensional nature of reference, and by taking care to explicitly distinguish the notion of meaning or of necessity that is in question. In this way we can see that nothing in Kripke's account counts against the existence of conceptual truths. It is just that conceptual truths are now split into two distinct varieties, an *a priori* and an *a posteriori* sort.[24]

4. Logical necessity and logical supervenience

The notion of logical necessity required in our definition of logical supervenience is precisely necessity as explicated above. A statement is logically necessary if and only if it is true in all possible worlds. Of course we have two varieties of logical necessity, depending on whether we evaluate truth in a possible world according to prior or posterior intensions.

This analysis explicates the logical necessity and possibility of a *statement* in terms of (a) the possibility of *worlds*, and (b) the intensions determined by the terms involved in the statement. I will not engage the vexed question of the ontological status of possible worlds.[25] For our purposes we only require that (1) any conceivable world is possible (I say more on conceivability below), and (2) any recombination of the fundamental qualities present in our world is possible (within certain limits; see Lewis 1986 and Armstrong 1990 for more on this).

Logical necessity comes to just the same thing as truth in virtue of meaning, given that the meaning of any term determines an intension[26] and that it is the intension of a term that determines truth in a possible world. (Of course we have to make parallel decisions concerning prior and posterior varieties of necessity and meaning.)

[24]The relation between the second and third considerations in this section—that is, between Quine's empirical revisability and Kripke's *a posteriori* necessity—is complex and interesting. As Kripke observes, the framework he develops accounts for some but not all of the problems raised by Quine. Kripke's analysis accounts for *a posteriori* revisions in posterior intensions, and therefore for changes in a certain sense of 'meaning'. However, the two-dimensional analysis agrees with the single-intension account on the truth-values it assigns at the actual world, so it does not account for the Quinean possibility of certain purported *a priori* conceptual truths turning out to be false in the actual world, in the face of sufficient empirical evidence. It seems to me that such purported conceptual truths are simply not conceptual truths at all, although they may be close approximations.

[25]Worlds should be seen prelinguistically, perhaps as distributions of basic qualities. Worlds should not be seen as collections of statements, as statements *describe* a world, and we have seen that they can do so in more than one way. To regard a world as a collection of statements would be to lose this distinction. Perhaps worlds can be regarded as collections of propositions (Adams 1974), if propositions are understood appropriately, or as maximal properties (Stalnaker 1976), or as states of affairs (Plantinga 1976), or as structural universals (Forrest 1986), or as concrete objects analogous to our own world (Lewis 1986a). Perhaps the notion can simply be taken as primitive. In any case, talk of possible worlds is as well- or poorly-grounded as talk of possibility and necessity in general. As with mathematical notions, these modal notions can be usefully deployed even preceding a satisfying ontological analysis. The intuitive understanding of a possible world as a world that God might have created suffices for most purposes (except perhaps for questions about God himself!).

If a statement is true in virtue of meaning, then the relevant intensions will ensure that the statement is true in all possible worlds. If the statement is not true in virtue of meaning, then the falsity of the statement will be compatible with the meanings involved and therefore with the intensions involved, so that there will be a possible world in which the statement is false. This allows us to freely move between the world-based and meaning-based notions of possibility, as long as we are careful to keep the intensions (prior or posterior) straight.

It also turns out that a statement is logically possible if it is conceivable (or conceivably true) in a specific sense. There is another sense of "conceivable" according to which a statement is conceivable if for all we know it is true, or if we do not know that it is impossible—in this sense, both Goldbach's conjecture and its negation is conceivable, even though one of them holds in no possible worlds. The sense of conceivability we need is rather: a statement is conceivable if it is true in some conceivable situation (or conceivable world). In this sense, the false member of the Goldbach pair does not qualify as a conceivable statement, as it is false in all conceivable situations. Given that every conceivable world is a possible world, the conclusion follows.

Our conceivability judgments are fallible, of course. We can go wrong in claiming that a statement is conceivable even if we have a situation in mind, because we can go wrong as describing the situation as one in which the statement is true.[27] However, given that we conceive a situation and describe it correctly by the statement in question, using the appropriate intensions for the terms in question, conceivability judgments will be reliable; we simply must be very careful to interpret a conceived situation correctly.

---

[26]Some have argued that in some cases—mathematical truths such as "there are an infinite number of primes", perhaps—statements can be necessary without being conceptually true. How can this be, given the argument above? I think there is an explanation: in these cases it is arguable that *truth-conditions* of a statement can be understood without clear intensions for the terms involved. The intensions associated with mathematical terms are unclear, precisely because the reference of those terms is unclear. What does "prime number" pick out in a world, for instance; for that matter what does "three" pick out? A Platonic object? Some complex physical entity? Nothing at all? Insofar as the meaning of the terms determines reference and intension, then the statements will be conceptually true, but it is arguable that the truth of mathematical statements can be understood even in the absence of referents for the terms involved; we do not need a world to evaluate the truth of a mathematical statement.

I think that mathematical truths can be understood as conceptual truths by analyzing intensions appropriately, but in any case the matter is unimportant. We will always be dealing with terms that pick out contingent natural phenomena in a world, and which therefore have reference and intension determined by meaning and playing a role in determining truth and truth-conditions. In such cases conceptual and necessary truth will coincide.

[27]For instance, one might claim that the falsity of the four-color theorem is conceivable, by imagining a situation in which the mathematical community proclaims that there is a counterexample. Nevertheless, this situation is not one in which the four-colour theorem is false, as the theorem is

An implication in the other direction, from logical possibility to conceivability, is trickier in that limits on our cognitive capacity imply that there is some possible situations that we cannot conceive, perhaps due to their great complexity. However, if we understand conceivability as conceivability-in-principle—perhaps conceivability by a superbeing— then it is plausible that logical possibility implies conceivability. In any case, we will be more concerned with the other implication.

On this understanding, if a statement is logically possible or necessary according to its prior intension, the possibility or necessity is knowable *a priori*. Modality is not epistemically inaccessible; the possibility of a statement is a function of the intensions involved and the space of possible worlds, both of which are epistemically accessible, and neither of which is dependent on *a posteriori* facts in this case. In the case of possibility and necessity according to a posterior intension, the possibility and necessity is knowable only *a posteriori*, as contingent facts go into the determination of the posterior intension.

We obtain two slightly different notions of logical supervenience depending on whether we use the prior or posterior brands of logical necessity. If 'gloop' has both a prior and a posterior intension associated with it, then gloopness may logically supervene on physical properties according to either the prior or the posterior intension of 'gloop'. Supervenience according to posterior intension—that is, supervenience with *a posteriori* necessity as the relevant modality—corresponds to what some call 'metaphysical supervenience'. We have now seen how this can be regarded as a variety of logical supervenience.[28]

I will discuss both the prior and posterior versions of logical supervenience in specific cases, but the former will be more central. Especially when considering questions about explanation, prior intensions are more important than posterior intensions. At the start of the explanatory process, as we have seen, we have only the prior intension to work with, and it is this intension that determines whether or not an explanation

true in all possible situations; it is merely a situation in which the mathematical community says it is false. See Yablo 1993 for more on conceivability as a guide to possibility.

Some other examples are due to Kripke (1972): we might think that it is conceivable that water is not H_2O but XYZ, by conceiving a situation in which XYZ is the clear liquid in oceans and lakes; but this situation is more correctly described as one in which XYZ is not water but merely watery stuff. At least this is so when we use posterior intensions for description; if we use prior intensions, it really is conceivable that water is XYZ. Thus we have prior and posterior senses of conceivability as it applies to statements; this seems a useful way to make sense of Kripke's examples.

[28]There is really only one kind of logical supervenience of *properties*, just as there is only one kind of logical necessity of *propositions*. However, we have seen that terms or concepts effectively determine two properties, one via a prior intension ('watery stuff') and the other via a posterior intension ('H_2O'). So for a given concept ('water'), there are two ways in which properties associated with that concept might supervene. I will sometimes talk loosely of the prior and posterior intensions associated with a property, and of the two ways in which a property might logically supervene.

is satisfactory. To explain water, for example, we have to explain things like its clarity, liquidity, and so on. The posterior intension ('H_2O') does not emerge until after an explanation is complete, and therefore does not itself determine a criterion for explanatory success. It is logical supervenience according to a prior intension that determines whether reductive explanation is possible. Where I do not specify otherwise, it is logical supervenience according to a prior intensions that I will generally be discussing.

If we choose one sort of intension—say, the prior intension—and stick with it, then we can see that various ways of formulating logical supervenience are equivalent. According to the definition in 2.1, B-properties are logically supervenient on A-properties if for any nomically possible situation X and any logically possible situation Y, if X and Y are indiscernible with respect to their A-facts, then all the B-facts true of X are true of Y. This is equivalent to the simpler: for any nomically possible situation X, the A-facts about X *entail* the particular B-facts about X (where "P entails Q" is understood as "it is logically impossible that P and not Q").[29]

Sticking to global supervenience, this means that B-properties logically supervene on A-facts if the B-facts about our world (and about nomically possible worlds) are entailed by the A-facts. Similarly, B-properties logically supervene on A-properties if there is no conceivable world which has the same A-properties as our world (or as a nomically possible world) but different B-properties. We can also say that logical supervenience holds if, given the totality of A-facts A^* and any B-fact B about our world (or a nomically possible world) W, "$A^*(W) \rightarrow B(W)$" is true in virtue of the meanings of the A-terms and the B-terms.

Finally, if B-properties are logically supervenient on A-properties according to prior intension, then in principle someone who knows all the A-facts about a (nomically possible) situation will be able to ascertain the B-facts about the situation from those facts alone, given that they possess the B-concepts in question. This follows from the fact that the logically necessary implication from A-facts to B-facts is knowable *a priori*. For logical supervenience according to posterior intension, B-facts about a situation can also be ascertained from the A-facts, but only *a posteriori*. The A-facts will have to be supplemented with contingent facts about the actual world,

[29]Note that entailment should not be understood in terms of the ability to derive one fact from another in a formal system. Such formal systems are generally inadequate reflections of meaning. A system such as the predicate calculus is constructed so that entailment within the system mirrors entailment based on the meaning of a few terms such as "and", "or", "some", and "all"; but it does not reflect the meaning of the vast majority of terms in our language. One could perhaps try to build in the meanings of those terms by axiomatization, but any such axiomatization is likely to be imperfect. In any case, the justification of the axioms is dependent on their being logically necessary—that is, true in virtue of the meanings of the concepts involved. The syntactic formalism is therefore derivative on these semantic notions.

as those facts will play a role in determining the B-intensions involved. Of course, this sort of inference may be quite impossible in practice, due to the complexity of the situations involved, but it is at least possible in principle.

There are therefore at least three avenues to establishing claims of logical supervenience: these involve conceivability, epistemology, and analysis. To establish that B-properties logically supervene on A-properties, we can (1) argue that instantiation of A-properties without instantiation of the B-properties is inconceivable; (2) argue that someone in possession of the A-facts could come to know the B-facts (at least in cases of supervenience via prior intension); (3) analyze the intensions of the B-properties in sufficient detail that it becomes clear that B-statements follow from A-statement in virtue of meaning alone. The same goes for establishing the failure of logical supervenience. I will use all three methods in arguing for central claims involving logical supervenience.

Not everybody may be convinced that the various formulations of logical supervenience are equivalent, so when arguing for important conclusions involving logical supervenience I will run versions of the arguments using each of the different formulations. In this way it will be seen that the arguments are robust, with nothing depending on a subtle equivocation between different notions of supervenience.

2.5 Almost everything is logically supervenient on the physical

In the following chapter I will argue that conscious experience does not logically supervene on the physical, and therefore cannot be reductively explained. A frequent response is that conscious experience is not alone here, and that all sorts of properties fail to logically supervene on the physical. It is suggested that such diverse properties as tablehood, life, and economic prosperity have no *logical* relationship to facts about atoms, electromagnetic fields, and so on. Surely those high-level facts could not be logically entailed by the microphysical facts?

On a careful analysis, I think that it is not hard to see that this position is wrong, and that the high-level facts in question (globally) logically supervene on the physical insofar as they are facts at all.[30] Conscious experience, and facts dependent on the existence of conscious experience, are almost *sui generis* in their failure to logically supervene. The relationship that consciousness bears to the physical facts is entirely different in kind to the standard sort of relationship between high-level and low-level facts.

[30]Horgan 1984c sets out and argues for the position that all high-level facts logically supervene on microphysical facts. As he puts it, those facts are tied to the microphysical by "semantic constraints",

There are various ways to make it clear that most properties logically supervene on physical properties. Here I will only be concerned with properties that characterize *natural phenomena*—that is, contingent aspects of the world that we have good reason to believe in and that need explaining. The property of being an angel might not logically supervene on the physical, but angels are not phenomena, so this failure need not concern us. I will also not concern myself with facts about abstract entities such as mathematical entities and propositions, which need to be treated quite separately. (See Armstrong 1982 for arguments that these logically supervene on the physical.)

Before proceeding, it is worth noting that in claiming that most high-level properties supervene on the physical, I am not arguing that high level facts and laws are entailed by microphysical *laws*, or even by microphysical laws in conjunction with boundary conditions. That would be a strong claim, and although it might have some plausibility if qualified appropriately, the evidence is not yet in. I am making the much weaker claim that high-level facts are entailed by all the microphysical *facts* (perhaps along with microphysical laws). This enormously comprehensive set includes the facts about the distribution of every last particle and field in every last corner of space-time: from the atoms in Napoleon's hat to the electromagnetic fields in the outer ring of Saturn. Fixing this set of facts leaves very little room for anything else to vary, as we shall see.

Before moving to the arguments I should note two harmless reasons why logical supervenience on the physical sometimes fails. First, some high-level properties fail to logical supervene because of an intrinsic dependence on conscious experience. Perhaps conscious experience is partly constitutive of a property like love; and as we will see, the prior (although not the posterior) intensions associated with some external properties such as color and heat may be dependent on phenomenal qualities. These should not be seen as providing counterexamples to my thesis, as they introduce no new failure of logical supervenience. Perhaps the best way to phrase the claim is to say that all facts logically supervene on the combination of physical facts and facts about conscious experience, or that all facts logically supervene on the physical facts *modulo conscious experience*. Similarly, a dependence on conscious experience may hinder the reductive explainability of some high-level phenomena, but we can still say that they are reductively explainable modulo conscious experience.

Second, an *indexical* element also enters into the application of some prior intensions, although not posterior intensions, as we saw in 2.4. The prior intension of 'water', for example, is something like 'the clear, drinkable liquid in our environment', so that if there is watery H_2O and watery XYZ in the actual world, which of them

---

so that all there is in the world is microphysics and "cosmic hermeneutics". He conspicuously avoids the problem of conscious experience, however. Others who advocate a version of the logical supervenience thesis include Jackson 1993, Kirk 1974, and Lewis 1994.

qualifies as 'water' depends on which is in the environment of the agent using the term. In principle we therefore need to add an *index* representing the location of an agent to the supervenience base in some cases. This gives us logical supervenience and reductive explanation modulo conscious experience and indexicality.

Finally, cases where the high-level facts are indeterminate do not count against logical supervenience. The claim is only that insofar as the high-level facts are determinate, they are determined by the physical facts. If the world itself does not suffice to fix the high-level facts, we cannot expect the physical facts to.[31]

I will argue for the ubiquity of logical supervenience using arguments that appeal to conceivability, to epistemological considerations, and to analysis of the notions involved.

1. *Conceivability.* The logical supervenience of most high-level facts is most easily seen by using conceivability as a test for logical possibility. What kind of world could be identical to ours in every last microphysical fact, but biologically distinct? Say a wombat has had two children in our world. The physical facts about our world will include facts about the distribution of every particle in the spatiotemporal hunk corresponding to the wombat, and its children, and their environments, and their evolutionary histories. If a world shared those physical facts with ours, but was not a world in which the wombat had two children, what could that difference consist in? Such a world seems quite inconceivable. Once a possible world is fixed to have all those physical facts the same, then the facts about wombat-hood and parenthood are automatically fixed. These biological facts are not the sort of thing that can float free of their physical underpinnings even as a conceptual possibility.

The same goes for architectural facts, astronomical facts, behavioral facts, chemical facts, economic facts, meteorological facts, sociological facts, and so on. A world physically identical to ours, but in which these sort of facts differ, is inconceivable. Furthermore this inconceivability does not seem to be due to any contingent limits in our cognitive capacity. Such a world is inconceivable *in principle*. Even a superbeing, or God, could not imagine such a world. There is simply not anything for them to imagine. Once they imagine a world with all the physical facts, they have automatically imagined a world in which all the high-level facts hold. Therefore a physically

[31] Some might suggest that logical supervenience would fail if there were two equally good high-level theories of the world which differ in their description of the high-level facts. One theory might hold that a virus is alive, for instance, whereas another might hold that it is not, so the facts about life are not determined by the physical facts. (The indeterminacy of translation provides another example.) This is not a counterexample, however. It is simply a case where the facts about life are indeterminate, so that we are free to legislate them one way or another according to what is useful. If the facts *are* determinate—e.g. if it is true that viruses are alive—then one of the descriptions is simply wrong. Either way, insofar as the facts about the situation are determinate at all, they are entailed by the physical facts.

identical world in which the high-level facts are false is logically impossible,[32] and the high-level properties in question are logically supervenient on the physical.

2. *Epistemology.* Moving beyond conceivability intuitions, we can note that if there *were* a possible world physically identical to ours but biologically distinct, then this would raise radical epistemological problems. How would we know that we were not in that world rather than this one? How would we know that the biological facts in our world are as they are? To see this, note that if I were in the alternative world, it would certainly *look* the same as this one. It instantiates the same distribution of particles found in the plants and animals in this world; indistinguishable patterns of photons are reflected from those entities; no difference would be revealed under even the closest examination. It follows that all the external evidence we possess fails to distinguish the possibilities. Insofar as the biological facts about our world fail to logically supervene, there is no way we can know those facts on the basis of external evidence. Given that there is no deep epistemological problem about biology, or economics, or architecture, it follows that facts in those domains logically supervene on the physical.

We can back up this point by noting that in areas where there *are* epistemological problems, there is an accompanying failure of logical supervenience, and that conversely, in areas where logical supervenience fails, there are epistemological problems.

Most obviously, there is an epistemological problem about consciousness, namely the problem of other minds. This problem arises because it seems logically compatible with all the external evidence that beings around us are conscious, and it is logically compatible that they are not. We have no way to peek inside a dog's brain, for instance, and observe the presence or absence of conscious experience. The status of this problem is controversial, but the mere *prima facie* existence of the problem is sufficient to defeat an epistemological argument, parallel to those above, for the logical supervenience of consciousness. By contrast, there is not even a *prima facie* problem of other biologies, or other economies. Those facts are straightforwardly publically accessible, precisely because they are fixed by the physical facts.[33]

Another famous epistemological problem concerns facts about causation. As Hume argued, external evidence only gives us access to regularities of succession between events; it does not give us access to any further fact of causation. So if

[33] Question: Why doesn't a similar argument force us to the conclusion that if conscious experience fails to logically supervene, then we can't know about even our *own* consciousness? Answer: Because conscious experience is at the very center of our epistemic universe. The skeptical problems about non-supervenient biological facts arise because we only have access to biological facts by external, physically mediated evidence; external non-supervenient facts would be out of our direct epistemic reach. There is no such problem with our own consciousness. I will discuss the epistemology of our own conscious experience further in chapter 6.

causation is construed as something over and above the presence of a regularity (as I will assume it must be), it is not clear that we can know it exists. Once again, this skeptical problem goes hand in hand with a failure of logical supervenience. In this case, facts about causation fail to logically supervene on matters of particular physical fact. Given all the facts about distribution of physical entities in space-time, it is logically possible that all the regularities therein arose as a giant cosmic coincidence without any real causation. At a smaller scale, given the particular facts about any apparent instance of causation, it is logically possible that it is a mere succession. We infer the existence of causation by a kind of inference to the best explanation—to believe otherwise would be to believe in vast, inexplicable coincidences—but belief in causation is not forced on us in the way that belief in biology is forced on us.

I have sidestepped problems about the supervenience of causation by stipulating that the supervenience base for our purposes includes not just particular physical facts but all the physical laws. It is reasonable to suppose that the addition of laws fixes the facts about causation. But of course there is a skeptical problem about laws paralleling the problem about causation: witness Hume's problem of induction, and the logical possibility that any apparent law might be an accidental regularity.

As far as I can tell, these two problems exhaust the epistemological problems that arise from failure of logical supervenience on the physical. There are some other epistemological problems that in a sense precede these, because they concern the existence of the physical facts themselves. First, there is Descartes' problem about the existence of the external world. It is compatible with our experiential evidence that the world we think we are seeing does not exist; perhaps we are hallucinating, or we are brains in vats. This problem can be seen to arise precisely because the facts about the external world do not logically supervene on the facts about our experience. (Idealists, positivists, and various others have argued controversially that they do. Note that if these views are accepted the skeptical problem falls away.) There is also an epistemological problem about the theoretical entities postulated by science—electrons, quarks, and such. Their absence would be logically compatible with the directly observable facts about objects in our environment, and some have therefore raised skeptical doubts about them. This problem can be analyzed as arising from the failure of theoretical facts to logically supervene on observational facts. In both these cases, skeptical doubts are perhaps best quelled by a form of inference to the best explanation, just as in the case of causation, but the in-principle possibility that we are wrong remains.

In any case, I am bypassing the latter skeptical problems by giving myself the physical world for free, and fixing all physical facts about the world in the supervenience base (therefore assuming that the external world exists, and that there are electrons, and so on). Given that those facts are known, there is no room for skeptical doubts about most high-level facts, precisely because they are logically supervenient.

To put the matter the other way around: all our sources of external evidence logically supervene on the microphysical facts, so that insofar as some phenomenon does not supervene on those facts, external evidence can give us no reason to believe in it. One might wonder whether some further phenomena might be posited via inference to the best explanation, as above, to explain the microphysical facts. Indeed, this process takes us from particular facts to simple underlying laws (and hence gives us causation, as we saw above), but then the process seems to stop. It is in the nature of fundamental laws that they are the end of the explanatory chain (except perhaps for theological speculation). This leaves phenomena that we have *internal* evidence for—namely conscious experience—and that is all. Modulo conscious experience, all phenomena must logically supervene on the physical.

We can make the case for logical supervenience by more direct epistemological considerations, arguing that someone in possession of all the physical facts could in principle come to know all the high-level facts, given that they possess the high-level concepts involved. Certainly one could never *in practice* ascertain the high-level facts from the set of microphysical facts. The vastness of the latter set is enough to rule that out. (Even less am I arguing that one could perform a derivation in any given formal system; for reasons canvassed earlier, formal systems are irrelevant here.) But as an in-principle point, there are various ways to see that someone (a superbeing?) armed with only the microphysical facts and the concepts involved could infer the high-level facts.

The simplest way is to note that in principle one could build a big mental simulation of the world and watch it in one's mind's eye, so to speak. Say that Joe is carrying an umbrella. From the associated microphysical facts, one could straightforwardly infer facts about the distribution and chemical composition of mass in Joe's vicinity, giving a high-level structural characterization of the area. One could determine the existence of a male fleshy biped straightforwardly enough. For instance, from the structural information one could note that there was an organism atop two longish legs that were responsible for its locomotion, that the creature has male genitalia, and so on. It would be clear that there he was carrying some contraption that was preventing drops of water, otherwise prevalent in the neighborhood, from hitting him. Doubts that this contraption is really an umbrella could be assuaged by noting from its physical structure that it can fold and unfold; from its history that it was hanging on a stand that morning, and was originally made in a factory with others of a similar kind, and so on. Doubts that the fleshy biped is really a human could be assuaged by noting the composition of his DNA, his evolutionary history and his relation to other beings, and so on. We need only assume that the being possesses enough of the concept involved to be able to apply it correctly to instances (that is, the being possesses the intension). If so, then the microphysical facts will give it all the evidence it needs to apply the concepts, and to determine that there really is a person carrying an umbrella here.

The same goes for almost any sort of high-level phenomena: tables, life, economic prosperity. By knowing all the low-level facts, a being in principle can infer all the facts necessary to determine whether or not this is an instance of the property involved. Effectively, what is happening is that a possible world compatible with the microphysical facts is constructed, and the high-level facts are simply read off that world using the appropriate intension (as the relevant facts are invariant across physically identical possible worlds). Hence the high-level facts logically supervene on the physical.

3. *Analyzability.* I have so far been inexplicit about the concepts involved in characterizing high-level phenomena. All I have needed is that insofar as those concepts have intensions allowing entities in the actual world to be picked out and facts about the actual world to be true, the same intensions should generally allow one to infer those facts from knowledge of the microphysical facts. This is perhaps the crux of the argument: insofar as there is truth in attributions of external facts, there is logical supervenience. It is nevertheless worth examining some specific concepts to note the nature of the intensions that make such inference possible.

There are some obstacles to elucidating these intensions, and to summarizing them in words. The first lies in the fact that as we saw earlier, application-conditions of a concept are often vague, being indeterminate in places. Is a cup-shaped object made of tissues a cup? Is a computer virus alive? Is a book that coagulates randomly into existence a book? Our ordinary concepts do not give straightforward answers to these questions; in a sense, it is a matter for stipulation. Hence there will not be determinate application-conditions for use in the entailment process. But as we saw in 2.4, this indeterminacy precisely mirrors an indeterminacy about the facts themselves. Insofar as the intension of 'cup' is a matter for stipulation, the facts about cups are also a matter for stipulation. What counts for our purposes is that the intension together with the microphysical facts determines the high-level facts insofar as they are really factual. Although vagueness and indeterminacy make discussion somewhat awkward, it affects nothing important to the issues.

A related problem is that any short analysis of a concept in terms of other concepts will invariably fail to do justice to the original concept. As we saw earlier, concepts do not usually have crisp definitions. At a first approximation, we can say something is a table if it has a flat horizontal surface with legs as support; but this lets in too many things (Frankenstein's monster on stilts?) and omits others (a table with no legs, sticking out from a wall?). One can refine the definition, adding further conditions and clauses, but we quickly hit the problems with indeterminacy, and in any case the product will never be perfect. All these complications are internal to the concept, however, and reflect no deep metaphysical facts. The ontological status of tablehood is not much different from that of schmablehood, where a schmable is defined to be something with a flat top and legs; the intension associated with tablehood is just less easy to characterize in detail. If schmablehood logically supervenes, so does tablehood. If one wanted, one could give a drawn-out analysis that approximated tablehood to near the point where indeterminacy creeps in, but there would be little point. Instead, to make the point about logical supervenience it suffices to gesture at a rough-and-ready analysis of the concepts involved. Any details would just be more of the same.

The point here is not to produce a perfect analysis, but to characterize a concept in a way that makes clear how it will allow entailment by microphysical properties. As we saw earlier, we do not need a definition of B-properties in terms of A-properties in order that A-facts entail B-facts. Meanings are fundamentally represented by intensions, not definitions; the role of analysis here is simply to characterize the intensions in sufficient detail that the entailment becomes clear. Intensions generally apply to individuals in a possible world in virtue of some of their properties and not in virtue of others; the point of this sort of analysis is to see what sort of properties the intension applies in virtue of, and to make the case that these are the sort of properties compatible with entailment.

A third problem lies in the fact that as we saw in 2.4, many concepts do not have just one set of application-conditions. Instead, there is a prior intension that fixes reference, and a posterior intension that determines reference in counterfactual situations. This is not much of a problem, as long as we keep the intensions separate. The posterior intension associated with 'water' is something like 'H_2O', which is obviously logically supervenient on the physical. But the prior intension, something like 'the clear, drinkable liquid in our environment' is equally logical supervenient, as the clarity, drinkability, and liquidity of water is entailed by the physical facts.[34] We can run things either way. As we have seen, it is the prior intension that enters into reductive explanation, so it is this that we are most concerned with. In general, if a prior intension I logically supervenes on the physical, then the posterior intension $dthat(I)$ will certainly logically supervene, as it will generally consist in a projection of some intrinsic physical structure across worlds.

Considerations about *a posteriori* necessity have led some to suppose that there

[34]Conscious experience arguably contributes to the prior intension, if water is individuated partly by the kind of experience it gives rise to. Indexicality certainly contributes, as seen by the 'our' in the 'the clear, drinkable liquid in our environment'. These facts do not undermine logical supervenience modulo conscious experience and indexicality.

As before, my summarizing prior intensions in short phrases like the above should not be seen as advocating "descriptive" theories of reference over "causal" theories. It is very likely that appropriate causal connections to the speaker are partly constitutive of the prior intension of a term like 'water'.

can be no logical entailment from low-level facts to high-level facts. Typically one hears something like "water is necessarily H_2O, but that is not a truth of meaning, so there is no meaning relation". But this is a vast oversimplification. For a start, the posterior intension 'H_2O' can be seen as part of the "meaning" of 'water' in some sense, and it certainly logically supervenes. But just as importantly, the prior intension ('the clear, drinkable liquid...') which fixes reference also supervenes, perhaps modulo experience and indexicality. It is precisely in virtue of its satisfying this intension that we deemed that H_2O was water in the first place. Given the prior intension I, the high-level facts are derivable unproblematically from the microphysical facts (modulo the contribution of experience and indexicality). The Kripkean observation that the concept is better represented as $dthat(I)$ affects this derivability not at all. The conceptual phenomenon of rigidification does not alone make an ontological difference.

With these obstacles out of the way, we can look at the intensions associated with various high-level concepts. It seems to me that most of these are characterizable in functional or structural terms, or as a combination of the two. For example, the sorts of things relevant to something's being a table include (1) that an object have a flat top and be supported by legs, and (2) that people use it to support various objects. The first of these is a structural condition: that is, a condition on the intrinsic physical structure of the object. It is clear that this sort of property is entailed by microphysical facts alone. The second is a functional condition: that is, it concerns the external causal role of an entity, characterizing the way it interacts with other entities. Functional properties are also generally logically supervenient. This is not quite as straightforward, because (a) such properties are dependent on a much wider supervenience base of microphysical facts, so one has to examine the facts about an object's environment; and (b) insofar as such properties are characterized dispositionally (something is soluble if it *would* dissolve *if* immersed in water), one needs to appeal to counterfactuals. But the truth-values of those counterfactuals are fixed by the inclusion of physical laws in the antecedent of our supervenience conditionals, so this is not a problem.

To take another example, the conditions on life roughly come down to some combination of the ability to reproduce, to adapt, and to metabolize, among other things (as usual, we need not legislate the weights, or all other relevant factors). These properties are all characterizable functionally, in terms of an entity's relation to other entities, its ability to convert external resources to energy, and its ability to react appropriately to its environment. These functional properties are all derivable, in principle, from the physical facts. As usual, there is no perfect *definition* of life in functional terms. The point is simply that this sort of characterization shows us that life is a functional property, whose instantiation can therefore be entailed by physical facts.

A complication is raised by the fact that functional properties are often characterized in terms of a causal role relative to other high-level entities. It follows that logical supervenience of the properties depends on the logical supervenience of the other high-level notions involved, where these notions may themselves be characterized functionally. This is ultimately not a problem, as long as causal roles are ultimately cashed out by non-functional properties: typically either by structural or phenomenal properties. There may be some circularity in the interdefinability of various functional properties—perhaps it is partly constitutive of a stapler that it deliver staples, and partly constitutive of staples that they are delivered by staplers—but this circularity can be handled by the familiar Ramsey-sentence method (see Lewis 1972), as long as the analyses are eventually grounded in structural or phenomenal properties. (The appeal to phenomenal properties may seem to count against logical supervenience on the physical, but see below. In any case, it is compatible with logical supervenience modulo conscious experience.)

Many properties are characterized relationally, in terms of some relation to an entity's environment. Usually such relations are ultimately causal, so that the properties in question are functional, but not always: witness the property of being on the same continent as a duck. Similarly, some properties are dependent on history (although these can usually be construed causally); perhaps it is partly constitutive of being an umbrella that something was designed to be used as an umbrella. Such properties pose no special problems for logical supervenience, as the relevant historical and environmental facts will themselves be fixed by the global physical facts.

Even a complex economic fact such as 'there was economic prosperity in the 1950's' is characterizable in mostly functional terms, and so can be seen to be entailed by the physical facts. A full analysis would be very complicated and would be made difficult by the vagueness of the notion of prosperity, but to get an idea how it might go, one can ask: why do we say that there *was* economic prosperity in the 1950's? At a first approximation, because there was high employment, people were able to purchase unusually large amounts of goods, there was low inflation, much development in housing, and so on. One is then lead to analyses of housing (the kind of place people sleep and eat in), of employment (organized labor for reward), and of monetary notions (presumably money will be roughly analyzable in terms of the systematic ability to exchange for other items, and its value will be analyzable in terms of how much one gets in exchange). All these analyses are ridiculously oversimplified, but the point is clear enough. These are generally functional notions of the sort whose satisfaction one can ascertain given the physical facts alone.

Many have been skeptical of the possibility of conceptual analysis. Often this has been for reasons that do not make any difference to the arguments I am making—because of indeterminacy in our concepts, for example, or because they lack crisp definitions. Sometimes it has been for deep reasons that I cannot go into here, such as those of Heidegger. In any case, if what I have said earlier in this chapter is correct, and if the physical facts about a possible world fix the high-level facts, we should *expect* to be able to analyze the intension of the high-level concept in question, at least to a good approximation, in order to see how its application can be determined by the physical facts. This is what I have tried to do in the examples given here. Other examples can be treated similarly. (A similar point about the requirement of analyzability for supervenience is made by Jackson 1993 and Lewis 1994.)

I am not for a moment advocating a program of trying to perform such analyses in general. Concepts are far too complex and unruly for this to do much good, and any explicit analysis is usually a pale shadow of the real thing. What counts is the general moral that most high-level concepts are not primitive, unanalyzable notions. They are generally analyzable as intensions specifying functional or structural properties. It is in virtue of this analyzability that high-level facts are in principle derivable from microphysical facts, and it is in virtue of this analyzability that high-level facts are reductively explainable by physical facts, at least in the weak sense set out above. By contrast, there seems to be no plausible functional, structural, or relational analysis of conscious experience, even of the most rough-and-ready kind. Consciousness seems only to be characterizable in terms of its brute phenomenal feel, which has to be taken as a kind of primitive notion.

Some problem cases

There are some sorts of properties that might be thought to provide particular difficulties for logical supervenience, and therefore for reductive explanation. I will examine a number of such candidates, paying particular attention to the question of whether the associated phenomena pose problems for reductive explanation analogous to the problems posed by consciousness. It seems to me that with a couple of possible exceptions, these do not pose any separate problem.

1. *Properties with phenomenal reference-fixation.* As discussed already, some concepts' prior intensions are dependent on some relation to conscious experience. An obvious example is redness, taken as a property of objects. On at least some accounts, the prior intension associated with redness requires that to be red, something must be the kind of thing that tends to cause red experiences under appropriate conditions.[35] So in its prior intension, redness does not logically supervene on the physical, although it supervenes modulo conscious experience. On the other hand, its posterior intension almost certainly supervenes. If it turns out that in the actual world, the sort of thing that tends to cause red experience is a certain surface reflectance, then redness can be identified *a posteriori* with that reflectance. This logically supervenes on the physical alone; so there can be red things even in worlds without conscious experience.

[35]An alternative plausible account holds that for something to be red, it must be the kind of thing

We saw earlier that failure of a prior intension to logically supervene is associated with a failure of reductive explanation. So does reductive explanation fail for redness? The answer is yes, in a weak sense. If redness is construed as the tendency to cause red experiences, then insofar as experience is not reductively explainable, neither is redness. But one can come close. One can note *that* a certain physical quality causes red experiences; and one can even in principle explain the causal relation between the quality and red-*judgments*. It is just the final step to experience that goes unexplained. In practice, our strictures on explanation are weak enough that this sort of thing counts. To explain a phenomenon to which reference is fixed by some phenomenal quality, we do not require an explanation of experience; otherwise we would wait a long time. Explanation modulo experience is good enough. (Nagel 1974 and Searle 1990 also discuss the manner in which the experiential quality in the explanation of external qualities is invariably omitted from what we take to be an explanation.)

A similar story holds for other phenomena, such as heat, light, and sound. The posterior intensions supervene on the physical facts, even if the prior intensions only supervene modulo conscious experience. In general, the posterior intension determines some structural property (molecular motion, the presence of photons, waves in air). Other phenomena may be characterized (in prior intension) by their relations to phenomena like heat and sound. As a reductive explanation here, we generally accept an account of the causal relation they bear to the structural properties in question. Again, this is an instance of reductive explanation modulo conscious experience.

2. *Consciousness-dependent properties.* Other properties have an even more direct dependence on conscious experience, such that experience not only plays a role in reference-fixation, but is partly constitutive of the *a posteriori* notion as well. The property 'stands next to a conscious person' is an obvious example. On some accounts, mental properties such as love and belief, although not themselves phenomenal properties, have a conceptual dependence on the existence of conscious experience. If so, then in a world without consciousness, such properties would not be exemplified. Therefore such properties fail to logically supervene even *a posteriori*, and reductive explanation fails even more strongly than in the above cases. But no *further* failure of explanation will arise, as they will be logically supervenient and reductively explainable modulo conscious experience.

3. *Intentionality.* It is worth separately considering the status of intentionality, as this is sometimes thought to pose problems analogous to those posed by consciousness. It seems to me that insofar as intentional properties fail to logical supervene,

---

that tends to cause red-*judgments*. This would eliminate the problems discussed here, as judgments are plausibly logically supervenient on the physical.

this is entirely derivative on the non-supervenience of consciousness. If intentional properties fail to logically supervene, this is only because some phenomenal quality is partly constitutive of intentional content. (Searle 1990 has argued for such a view of intentionality.) I think this is not the best account of intentionality, and that intentional content should be best analyzed in terms of causal relations that internal states bear to behavior and the environment, but as in Chapter 1 there is no need to legislate the issue. We can simply note that there is no third sense, so there is no reason to believe that intentionality fails to supervene in a non-derivative way.

Another way to see this, recalling our earlier discussion of epistemology, is to ask: why do we believe in intentionality at all? Presumably these reasons come to at most (a) internal evidence—our direct experience of it; and (b) external evidence—that is, as an inference from human behavior. The first kind is evidence about consciousness, and the second is evidence about behavior. Neither of these gives any reason to believe in a non-functional, non-phenomenal notion of intentionality.

Leaving the phenomenal notion aside, intentional properties are best seen as a kind of theoretical construct in the explanation of human behavior, and should therefore be analyzable in terms of causal connections to behavior and the environment. If so, then intentional properties are straightforwardly logically supervenient on the physical. Lewis (1974) makes a thorough attempt at explicating the entailment from physical facts to intentional facts by giving an appropriate functional analysis. More recent accounts of intentionality, such as those by Dennett (1987), Dretske (1981), and Fodor (1987) can be seen as contributing to the same project. The details of the required analysis are far from clear, but arriving at the details is a problem of explication rather than of explanation. There is no separate *ontological* problem of intentionality analogous to that of consciousness.

4. *Moral and aesthetic properties.* Moral properties and aesthetic properties are often held not to logically supervene on the physical. According to Moore (1922), there is no way of deriving an 'ought' from an 'is'. Nothing about the *meaning* of notions like 'goodness' allows that facts about goodness should be entailed by physical facts. Does this therefore make moral properties analogous to conscious experience?[36]

There is a strong disanalogy. Moral facts are not natural phenomena that force themselves on us. When it comes to the crunch, we can deny that they exist at all. Indeed, this precisely reflects the strategy taken by moral anti-realists such as Hare (1952) and Blackburn (1971), who argue that precisely because moral properties fail to

[36]One might be tempted to answer Moore by saying that moral properties supervene on the physical modulo conscious experience. After all, it is not entirely implausible that conscious experience is partly constitutive of morality. But this will not help. If Moore's argument succeeds, a modified version will equally show that moral facts are not entailed by the combination of physical facts and phenomenal facts.

logically supervene, they have no objective existence and instead should be relativized away as constructs or projections of our cognitive apparatus. This strategy cannot be taken for phenomenal properties, which undeniably exist (some have denied their existence, of course, but the denials have not been convincing).

It strikes me that there are about four alternatives. Moral properties might (a) logically supervene according to their prior intension; (b) failing that, they might supervene according to their posterior intensions; (c) failing that, they might supervene by a fundamental irreducible link; or (d) they have no objective existence. Moore argued against (a) with his "open question" argument, and instead advocated moral non-naturalism, a version of (c). The consequent "queerness" of moral properties has proved unpopular, however, as we have little reason to believe in such metaphysically independent facts (see e.g. Mackie 1977). Some contemporary moral realists (e.g. Boyd 1988; Brink 1989) have argued for something like (b), holding that there is an *a posteriori* link even though there is no *a priori* link, typically making analogies between moral terms and terms like 'water'. This position might seem difficult to maintain, given that even *a posteriori* equivalences must be grounded in *a priori* reference-fixation; see Horgan and Timmons (1992a; 1992b) for a critique. As far as I can tell, either there is some kind of semantic relation between physical and moral properties (contra Moore), or there are no objective moral facts[37] (essentially because if the meaning of a moral term like 'good' determines a prior intension, then such a semantic relation will hold, and if the meaning does not determine a prior intension, then there is no reason to believe that reference is fixed at all). The same goes for aesthetic properties.

A good way to see the disanalogy between moral and phenomenal properties is to note that in the moral domain, the *phenomena* requiring explanation are our *ascriptions* of moral and aesthetic properties to individuals and objects. With consciousness, such ascriptions require explanation, but they are not *all* that requires explanation. Also requiring explanation is the phenomenon of consciousness itself.

5. *Names.* On some accounts, there is no analysis associated with a name such as 'Rolf Harris', which simply picks out its referent directly (e.g. the theory of Kaplan 1989). Does this mean that the property of being Rolf Harris fails to logically supervene on the physical? It seems to me that the posterior intension certainly logically

[37] If there are no objective moral facts, it is still possible that there could be some sort of *subjective* moral facts. On this view, moral attributions would have determinate truth-conditions, but these would be dependent on the ascriber. In this case, there would be logical supervenience modulo indexicality. This corresponds to a view sometimes called "subjectivist moral realism", where 'good' is interpreted as meaning something like 'good-for-me' (see Sayre-McCord 1989). The subjectivity involved—it turns out that two people arguing over what is 'good' might not be disagreeing at all—makes this a very weak kind of realism, however.

supervenes (e.g., Rolf might be the person conceived from such-and-such sperm and egg in all possible worlds). It seems to me also that every use of the name has some kind of prior intension attached—for me, it is something like 'the guy called "Rolf Harris" who bangs around on paint cans, and who bears the appropriate causal relation to me'—and this intension logically supervenes. Rather than justifying this, however, I can simply note that any failure to logically supervene will not be accompanied by an explanatory mystery. The property of being Rolf Harris does not constitute a phenomenon in need of explanation, as opposed to explication. What needs explaining is why there is a person named 'Rolf Harris', who bangs around on paint cans, and so on. These properties certainly logically supervene, and so are explainable in principle in the usual way.

6. *Indexicals.* As we have seen, reference-fixation of many concepts, from 'water' to 'my dog', includes an indexical element. Reference of these notions is fixed on the basis of both physical facts and an agent-relative "indexical fact" representing the location of an agent using the term in question. Such a fact is determinate for any given agent (e.g., me), so reference-fixation is determinate. Supervenience and explanation succeed modulo that indexical fact.

Does indexicality pose a problem for reductive explanation? For arbitrary speakers, perhaps not, as the "fact" in question can be relativized away. But for myself, it is not so easy. The indexical fact expresses something very salient about the world as I find it: that David Chalmers is *me*. How could one explain this seemingly brute fact? Indeed, is there really a fact here to be explained, as opposed to a tautology? The issue is extraordinarily difficult to get a grip on, but it seems to me that even if the indexical is not an objective fact about the world, it is a fact about the world as I find it, and it is the world as I find it that needs explanation. The nature of the brute indexical is quite obscure, though, and it is most unclear how one might explain it.[38]

It is tempting to look to consciousness. But while an explanation of consciousness might yield an explanation of "points of view" in general, it is hard to see how it could explain why a seemingly arbitrary one of those points of view is *mine*, unless solipsism is true. The indexical fact may have to be taken as primitive. If so, then we have a failure of reductive explanation distinct from and analogous to the failure with consciousness. Still, the failure is less worrying than that with consciousness, as the unexplained fact is so "thin" by comparison to the facts about consciousness in all its glory. Admitting this primitive indexical fact would not require nearly as much of a revision to our materialist worldview as admitting irreducible facts about conscious experience.

7. *Universally quantified facts.* As we saw in 2.1, facts whose expression requires

[38]See Nagel 1986, Chapter 4, for a provocative discussion of this matter.

universal quantifiers are not logically determined by the physical facts, or indeed by any set of particular facts. Consider such universally quantified facts about our world as: there are no angels; Don Bradman is the greatest cricketer; everything alive is based on DNA. All these could be falsified, consistently with all the physical facts about our world, simply by the addition of some new non-physical stuff: cricket-playing angels made of ectoplasm, for instance. Even addition of facts about conscious experience or indexicality cannot help here.[39]

Does this mean such universal facts are not reductively explainable? It seems so, insofar as there is no physical explanation of why there is no extra non-physical stuff in our world. That is indeed a further fact. The best way to deal with this situation is to introduce a second-order fact that says of the set of basic particular facts, be they microphysical, phenomenal, indexical, or whatever: *that's all*. This fact says that all the particular facts about the world are included in or entailed by the given set of facts. From this second-order fact, in conjunction with all the basic particular facts, all the universally quantified facts will follow.

Obviously, this does not constitute a very serious failure of reductive explanation. Presumably there will be such a "that's all" fact true of any world, and such a fact will never be entailed by the particular facts. It simply expresses the bounded nature of our world, or of any world. It is a cheap way to bring all the universally quantified facts within our grasp.

8. *Physical laws and causation.* On the most plausible accounts of physical laws, these fail to logically supervene on the physical facts, taken as a collection of particular facts about a world's spatiotemporal history. One can see this by noting the logical possibility of a world physically indiscernible from ours over its entire spatiotemporal history, but with different laws. For instance, it might be a law of that world that whenever two hundred tons of pure gold is assembled in a vacuum, it will transmute into lead. Otherwise its laws are identical, with minor modifications where necessary. As it happens, in the spatiotemporal history of our world, twenty tons of gold is never assembled in a vacuum, so that our world and this world have identical histories. Their laws nevertheless differ.

Arguments like this demonstrate that the laws of nature do not logically supervene on the collection of particular physical facts.[40] By similar arguments (as we discussed earlier) one can see that a causal connection between two events is something over and above a regularity between the events. (This is a controversial matter. Holders

[39]How do these facts evade the arguments for logical supervenience above? The argument from conceivability fails, as the angel example shows. The argument from epistemology fails, as there clearly *is* an epistemological problem about how we could know quantified facts of unrestricted scope (we cannot *know* that there are no angels). The argument from analyzability fails, as there is no analysis of universally quantified facts in particular terms.

of various "Humean" views argue that there is no such further fact associated with laws or causation, but it seems to me that they have the worse of the arguments here. Humean views of laws and causation can be found in Lewis 1986b, Mackie 1974, and Skyrms 1980. For arguments against such views, see Armstrong 1982, Carroll 1990, Dretske 1977, Molnar 1969, and Tooley 1977.)

I have bypassed these problems elsewhere by including physical laws in the supervenience base, but this steps over a metaphysical puzzle rather than answering it. It is true that these do not constitute as significant a failure of reductive explanation as consciousness does. The laws and causal relations are themselves posited to explain existing physical phenomena, namely the manifold regularities present in nature, whereas consciousness, by contrast, is a brute explanandum. Nevertheless the very existence of such further facts over and above the spatiotemporal manifold raises deep questions about their metaphysical nature. Apart from conscious experience (and perhaps indexicality), these constitute the only such further facts in which we have any reason to believe. It is natural to speculate that these two non-supervenient kinds, consciousness and causation, have a close metaphysical relation. I will pursue speculation along these lines in further work (summarized at the end of this work.)

Recap

The position we are left with is that almost all facts logically supervene on the physical facts (including physical laws), with possible exceptions for conscious experience, indexicality, and universally quantified facts. To put the matter differently, we can say that the facts about the world are exhausted by (1) particular physical facts, (2) laws of nature, (3) facts about conscious experience, (4) a second-order "that's all" fact, and perhaps (5) an indexical fact about my location.[41] (The last two are minor compared to the others, and the status of the last is dubious, but I have put them in for completeness.) Modulo conscious experience and indexicality, it seems that all particular facts are logically supervenient on the physical. Of course, to establish this conclusion conclusively would require a detailed examination of all kinds of phenomena of a kind that I cannot undertake here. But from what we have seen, it is *prima facie* reasonable to suppose that almost everything is logically supervenient in this way.

[40]How do laws evade the arguments for logical supervenience above? The argument from conceivability fails, as the example above shows. The argument from epistemology fails, as there clearly are problems with the epistemology of laws and causation, as Hume has shown us. The argument from analysis fails, as lawhood requires a counterfactual-supporting universal regularity, and counterfactuals cannot be analyzed in terms of particular facts about a world-history (*pace* Lewis 1973). The particular facts about the world's spatiotemporal history are compatible with the truth of all sorts of different counterfactuals.

(Creation myth: Creating the world, all God had to do was fix the facts above. For maximum economy of effort, he first fixed the laws of nature— i.e. the laws of physics and any laws relating physics to conscious experience. Next, he fixed the boundary conditions: perhaps a time-slice of physical facts, and maybe the values in a random-number generator. These combined with the laws to fixing the remaining physical and phenomenal facts. Last, he decreed: that's all. Interestingly, it seems beyond God's powers to fix the indexical fact. That fact is irreducibly subjective—it is only a fact for *me*. Perhaps this is another reason to be skeptical about it.

Epistemological myth: At first, I have only facts about my conscious experience. From here, I infer facts about middle-size objects in the world, and eventually micro-physical facts. From regularities in these facts, I infer physical laws, and therefore further physical facts. From regularities between my conscious experience and physical facts, I infer psychophysical laws, and therefore facts about conscious experience in others. I seem to have taken the abductive process as far as it can go, so I hypothesize: that's all. The world is much larger than it once seemed, so I single out the original conscious experiences as *mine*. Note the very different order involved here. One could almost say that epistemology recapitulates ontology backwards.)

The logical supervenience of most high-level phenomena is a conclusion that has not been as widely accepted as it might have been, even among those who discuss supervenience. Although the matter is often not discussed, many have been wary about invoking the *logical* modality as relevant to supervenience relations. As far as I can tell there have been a number of separate reasons for this hesitancy, all of which are ultimately irrelevant.

First, the problem with logically possible physically identical worlds with extra non-physical stuff (angels, ectoplasm) has led some to suppose that supervenience relations cannot be logical (Haugeland 1982; Petrie 1987); but we have seen how to fix this problem. Second, many have supposed that considerations about *a posteriori* necessity demonstrate that supervenience relations cannot be underwritten by meanings (Brink 1989; Teller 1984); but we have seen that supervenience relations based on *a posteriori* necessity can be seen as a variety of logical supervenience. Third, there is a general skepticism about the notion of conceptual truth, deriving from Quine; but we have seen that this is a red herring here. Fourth, worries about "reducibility" have led some to suppose that supervenience is not generally a conceptual relation (Hellman and Thompson 1975); but it is unclear that there are any good arguments against reducibility that are also good arguments against logical supervenience. Fifth, the very phenomenon of conscious experience is sometimes invoked to demonstrate that supervenience relations cannot be logical in general (Seager 1988); but we have seen that conscious experience is almost *sui generis* in its failure to logically supervene.

Finally, sometimes the fact that supervenience relations are not usually logical is simply stated without argument, presumably as something that any reasonable person must believe (Bacon 1985; Heil 1992).[42]

It seems to me that every supervenience relation of some high-level property upon the physical is either (1) a logical supervenience relation, of either the prior or posterior varieties, or (2) a contingent nomic supervenience relation. If neither of these hold for some apparent supervenience relation, then we have good reason to believe that there are no objective high-level facts of the kind in question (as, perhaps, for moral facts). I will argue further in Chapter 4 that there is no intermediate variety of supervenience between the logical and the nomic.

This provides a unified explanatory picture, in principle. Almost every phenomenon is reductively explainable, in the weak sense outlined earlier, except for conscious experience and perhaps indexicality, along with the rockbottom microphysical facts and laws, which have to be taken as fundamental.

It is worth taking a moment to answer a query due to Blackburn (1985): how do we explain the supervenience relations themselves? For a logical supervenience relation based on the prior intension of a concept, this is a simple matter of giving an appropriate analysis of the concept, perhaps in functional or structural terms, and noting that its reference is invariant across physically identical worlds. Here, the supervenience relation is itself an *a priori* conceptual truth. For a logical supervenience relation based on a posterior intension, the supervenience can be explained by noting how the prior intension of the concept picks out some actual-world reference class which is them projected (by rigidification) invariantly across physically identical worlds. All we need here for an explanation is an *a priori* conceptual analysis combined with contingent facts about the actual world (see Horgan and Timmons 1992b for more on this). On the other hand, a mere nomic supervenience relation will itself be a contingent law. At best it will be explainable in terms of more fundamental laws. At worst, the supervenience law will itself be fundamental. In either case, one explains certain regularities in the world by invoking fundamental laws, just as one does in physics, and as always, fundamental laws are where explanation must stop. Nomic supervenience is ontologically expensive, as we have seen, so it is fortunate that logical supervenience is the rule and nomic supervenience the exception.

[42]By contrast, those who appear to hold that logical supervenience is the rule rather than the exception include Armstrong 1982, Horgan 1984c, Jackson 1993, Lewis 1994, and Nagel 1974.

Chapter 3

Can Consciousness be Reductively Explained?

3.1 Does consciousness logically supervene?

In the last chapter, we saw that almost every natural phenomenon logically supervenes on the physical, and is therefore susceptible to reductive explanation. In what follows, I will argue that conscious experience escapes this reductive net: conscious experience does not logically supervene on the physical and therefore cannot be reductively explained.

In principle, we need to show that consciousness does not supervene globally—that is, that the totality of microphysical facts about the world does not entail the facts about consciousness. In practice, however, it is easier to run the argument in terms of local supervenience, arguing that the microphysical properties of an individual do not entail the facts about consciousness in that individual. It does not matter which way we run the argument; as we saw earlier, when it comes to consciousness, local and global supervenience stand and fall together. Insofar as consciousness supervenes on the physical at all, it almost certainly supervenes on the local physical properties of an individual. If one finds this disputable, then all the arguments can be run at the global level with appropriate alterations.

To establish the failure of logical supervenience, then, we will be largely concerned with the logical possibility of a *zombie*: someone or something physically identical to me (or to any other conscious being), but lacking conscious experiences altogether. At the global level, we will be concerned to establish the logical possibility of a *zombie world*: a world physically identical to ours, but in which there are no conscious experiences at all. In such a world, everybody is a zombie.

Arguing for a logical possibility is not completely straightforward. How, for instance, would one argue that a mile-high unicycle is logically possible? It just seems obvious. Although no such thing exists in the real world, the description certainly appears to be coherent. If someone objects that it is not logically possible—it merely seems that way— there is little we can say, except to repeat the description and assert that it is *obviously* coherent, and that there is no hidden contradiction lurking within.

I should confess at the outset that the logical possibility of zombies seems equally obvious to me. A zombie is just something physically identical to me, but which has no conscious experience—all is dark inside. While this is probably empirically impossible, it certainly seems that a coherent situation is described; I can discern no contradiction in the description. In some ways an assertion of this logical possibility comes down to table-pounding, but no more so than with the unicycle. Almost everybody, it seems to me, is capable of conceiving of this possibility. Some may be led to deny the possibility in order to make some theory come out right, but the justification of such theories should ride on the question of possibility, rather than the other way around.

In general, a certain burden of proof lies on those who claim that a given description is logically *impossible*. If they truly believe that a mile-high unicycle is logically impossible, they must give us some idea of where a contradiction lies, whether explicit or implicit. If they cannot point out something about the intensions of the terms 'mile-high' and 'unicycle' that might lead to a contradiction, then their case will not be convincing. On the other hand, it is no more convincing to give an obviously false analysis of the notions in question—to assert, for example, that for something to qualify as a unicycle it must be shorter than the Statue of Liberty. If no reasonable analysis of the terms in question points toward a contradiction, or even makes the existence of a contradiction plausible, then there is a natural assumption in favor of logical possibility.

That being said, there are some positive things that proponents of logical possibility can do to bolster their case. They can exhibit various indirect arguments, appealing to what we know about the phenomena in question and the way we think about hypothetical cases involving these phenomena, in order to establish that the obvious logical possibility really is a logical possibility, and really is obvious. One might spin a fantasy about an ordinary person riding a unicycle, when suddenly the whole system expands 1000-fold. Or one might describe a series of unicycles, each bigger than the last. In a sense, these are all appeals to intuition, and an opponent who wishes to deny the possibility can in each case assert that our intuitions have misled us, but the very obviousness of what we are describing works in our favor, and helps shift the burden of proof even further onto the other side.

In arguing that almost everything logically supervenes on the physical in Chapter 2, I put forward three different kinds of argument: arguments from conceivability, epistemology, and analyzability. In arguing *against* the logical supervenience of consciousness on the physical, arguments again fall into one of these three classes. I will give five arguments below. The first two arguments are essentially arguments from conceivability, the third and fourth proceed from epistemological considerations, and the fifth is an argument from analyzability. Most of these are direct counterparts of arguments put forward in Chapter 2.

Argument 1: From the conceivability of zombies

I have already given the first argument against the logical supervenience of consciousness. This argument is an appeal to the conceivability of a zombie, or of a zombie world. If a zombie world is conceivable in an appropriate sense (i.e., one can't just be misled into *thinking* that it is conceivable, for instance by conceiving of a world and *misdescribing* it as a zombie world), then it is logically possible, as we saw in chapter 2. If so, then consciousness fails to logically supervene.

It is worthwhile clarifying just what is being described. The subject under consideration is my zombie twin, who is molecule-for-molecule identical to me, and indeed identical in all the low-level properties postulated by a completed physics, but who lacks conscious experience entirely. To fix ideas, we can imagine that right now I am gazing out the window, experiencing some nice green sensations from seeing the trees outside, having some pleasant taste experiences through munching on a chocolate bar, and feeling a dull aching sensation in my right shoulder. My zombie twin is physically identical, and we may as well suppose that he is embedded in an identical environment.

My zombie twin will certainly be identical to me *functionally*: he will be processing the same sort of information, reacting in a similar way to inputs, with his internal configurations being modified appropriately and with indistinguishable behavior resulting. He will be *psychologically* identical to me, in the sense developed in Chapter 1. He will be perceiving the trees outside, in the functional sense, and having a pain, in the psychological sense. All of this follows logically from the fact that he is physically identical to me, by virtue of the functional analyses of psychological notions. It even follows that he will be "conscious" in the functional senses outlined in section 1.5—he will be awake, able to report the contents of his internal states, able to focus attention in various places, and so on. It is just that none of this functioning will be accompanied by any real conscious experience. There will be no phenomenal feel. There is nothing it is like to be a zombie.

Note that this sort of zombie is quite unlike the zombies found in Hollywood movies, which tend to be significantly functionally impaired. Insofar as they lack consciousness, they lack it in a psychological sense thereof: typically, they are quite lacking in introspective ability, and in the ability to voluntarily control behavior. They may or may not lack phenomenal consciousness; as Block (1993) points out, it is reasonable to suppose that there is something it tastes like when they eat their victims. We can call these *psychological zombies*; I am concerned with *phenomenal zombies*, which are functionally identical, but which lack experience.[1] (Perhaps it is not surprising that phenomenal zombies have not been popular in Hollywood, as there would be obvious problems with their depiction.)

---

[1]The psychological zombie/phenomenal zombie distinction corresponds to what Dennett (1991) calls the zombie/zimbo distinction.

Zombies as I have described them are a strange idea, and seem quite implausible as an empirical possibility. In practice, any replica of me would almost certainly have conscious experience. But the question is not whether a zombie replica is plausible; the question is whether it is a coherent notion. It certainly seems to be coherent, and this is sufficient to establish the conclusion.

As I said before, this argument inevitably rests on an appeal to intuition, but the intuition can be buttressed. A good way to buttress this intuition is an appeal to *bizarre realizations* of my functional organization.

My functional organization—that is, the pattern of causal organization embodied in the mechanisms responsible for the production of my behavior— can in principle be realized in all sorts of strange ways. To use a common example (Block 1978), the people of some large nation such as China might organize themselves so that they realize a causal organization isomorphic to that of my brain, with every person simulating the behavior of a single neuron, and with radio links corresponding to synapses. The population might control an empty shell of a robot body, equipped with sensory transducers and motor effectors.

Many people find it extremely implausible that a set-up like this would be accompanied by conscious experience—that somehow a "group mind" would emerge from the overall system. I am not concerned here with whether or not conscious experience would *in fact* arise; in fact, I think that it would, as I will argue in Chapter 7. All we need here is that the notion that such a system lacks conscious experience is *coherent*. A meaningful possibility is being expressed, and it is an open question whether consciousness arises or not. From this it follows that the existence of my conscious experience is not logically entailed by the facts about my functional organization.

But it is even clearer that there can be no logical entailment from the non-functional details of my physical structure to consciousness. There is certainly no more of a *conceptual* entailment from biochemistry to consciousness than there is from silicon, or from Chinese homunculi. If it is logically possible that the Chinese system lacks consciousness, it is equally possible that my replica lacks consciousness. Consciousness therefore fails to logically supervene on the physical.

Incidentally, I think it is no less clear that my replica could exist without consciousness than it is that the Chinese set-up could. The latter example is useful, however, in getting away from the very familiar case of our own biochemistry, wherein the constant conjunction with consciousness might lead one to mistakenly suppose a conceptual connection.[2]

[2]Jacoby (1990) makes the excellent point that conceivability arguments pose no more of a problem for functionalist accounts of consciousness than they do for materialist accounts in general. He takes this to be an argument for functionalist accounts, however, whereas I take it to be an argument against materialist accounts.

Argument 2: The Inverted Spectrum

The following argument does not strictly establish the logical possibility of zombies, but it is nevertheless strong enough to establish that consciousness does not logically supervene on the physical. One can coherently imagine somebody—my inverted twin—who is physically identical to me, but who differs in that whenever I have a red experience, he has a blue experience, and vice versa. He will call his blue experiences "red", of course, but that is irrelevant. What matters is that the experience he has of things we both call "red"—blood, fire engines, and so on—is of the same kind as the experience I have of the things we both call "blue", such as the sea and the sky.

The rest of his color experiences are systematically inverted with respect to mine, in order that they cohere with the red-blue inversion. Perhaps the best way to imagine this happening with human color experiences is to imagine that two of the axes of our three-dimensional color space are switched—the red–green axis is mapped onto the yellow–blue axis, and vice versa.[3] For such an inversion to happen in practice, one would presumably need to rewire the relevant neural processes in an appropriate way. But as a logical possibility, it seems entirely coherent that experiences could be inverted while physical structure is duplicated exactly. Nothing in the neurophysiology logically dictates that one sort of processing should be accompanied by red experiences rather than by yellow experiences.

It is sometimes objected that human color-space is asymmetrical in a way that disallows such an inversion (see Harrison 1973 and Hardin 1987). For instance, certain colors have a warmth or coolness associated with them that appears to make a functional difference; swapping a warm color and a cool color would cause the phenomenal feel to become dissociated from its accompanying functional role. There are three things we can say in response to this. First, there does not seem to be anything incoherent about the notion of such dissociation (e.g., warm phenomenology with cool reactions), although it is admittedly an odd idea. Second, instead mapping red precisely onto blue and vice versa, one can imagine that these are mapped onto slightly different colors—for example, red might be mapped onto a "warm" version of blue (see Levine 1991 for a development of this idea), or even onto some color not in our color space at all. Third, perhaps the most compelling response is to argue that even if our own color space is asymmetrical, one can certainly imagine creatures whose color space is symmetrical, and we can simply perform the inversion thought-experiment on them. For example, one might imagine a creature who sees precisely two colors, A and B, corresponding to distinct, well-separated ranges of light

[3]Actually, this will end up swapping red with yellow rather than blue, as both of these are at the positive ends of their axis. The details are inessential, however. See Hardin 1988 for a lucid discussion of the intricacies of human color-space.

wavelengths. It seems entirely coherent to imagine that there could be two physically identical creatures whose experiences of A and B were inverted.

Even many reductive materialists (e.g., Shoemaker 1982) have conceded that it is coherent that one's color experiences might be inverted while one's functional organization stays constant. Once this is granted, it follows that inversion of experiences in a physical replica is equally coherent. The extra neurophysiological properties that are constrained in such a case are not the kind of thing that could logically determine the nature of the experience.

While the possibility of inverted spectra and the possibility of zombies both establish that consciousness fails to logically supervene, the first establishes a conclusion strictly weaker than the second. Somebody might conceivably hold that inverted spectra but not zombies are logically possible. If this were the case, then the *existence* of consciousness could be reductively explained, but the specific *nature* of particular conscious experiences could not be.

Argument 3: From epistemic asymmetry

One of the most basic characteristics of consciousness, as we saw earlier, is its surprisingness. Our grounds for belief in consciousness derive solely from our own experience of it. Even if we knew every last detail about the physics of the universe—the configuration of, causation between, and evolution among all the fields and particles in the spatiotemporal manifold—it would nevertheless be the case that if we had not experienced consciousness for ourselves, there would be no reason to postulate the existence of such a phenomenon. My knowledge of consciousness, in the first instance, comes from my own case, not from any external observation. It is my own experience of consciousness that forces the problem on me.

From all the low-level facts about physical configurations and causation, we can in principle derive all sorts of high-level facts about macroscopic systems, their organization, and the causation among them. One could determine all the facts about biological function, and about human behavior and the brain mechanisms by which it is caused. But nothing in this vast causal story would ever lead one who had not experienced it directly to believe that there should be any *consciousness*. The very idea would be unreasonable; almost mystical, perhaps.

It is true that the physical facts about the world might provide some indirect evidence for the existence of consciousness. For example, from these facts one could ascertain that there were a lot of organisms that *claimed* to be conscious, and said they had mysterious subjective experiences. Still, this evidence would be quite inconclusive, and it would be most natural to draw an eliminativist conclusion—that is, that there was in fact no *experience* going on in these creatures, just a whole lot of talk.

Eliminativism about consciousness is an unreasonable position *only* because of our own experience of it. If it were not for this direct knowledge, consciousness could go the way of the vital spirit. To put it another way, there is an *epistemic asymmetry* in our knowledge of consciousness that is not present in our knowledge of other phenomena. Our knowledge that conscious experience exists derives solely from our own case, with external evidence being weak at best.

This way of looking at things makes it clear that consciousness cannot logically supervene on the physical. There can be no such epistemic asymmetry for logically supervenient facts, as we saw in Chapter 2. A logically supervenient property can be detected straightforwardly on the basis of external evidence, and there is no special role for the first-person case. (To be sure, there are some supervenient properties, such as belief, that are more easily detected in the first-person case. But this is just a matter of how hard one has to work. Beliefs are just as accessible in the third person, in principle; eliminativism about belief is made just as implausible by the third-person evidence as by the first-person evidence.) From the surprisingness and the epistemic asymmetry associated with consciousness, it follows that it cannot logically supervene on the physical. No collection of facts about complex causation in physical systems adds up to a fact about consciousness.

Argument 4: The Knowledge Argument

This epistemological argument is due to Jackson (1982), following a related argument by Nagel (1974). My version is slightly different from Jackson's, but the difference is inessential.

Imagine that we are living in an age of a completed neuroscience, where we know everything there is to know about the physical processes within our brain responsible for the generation of our behavior. Mary is colorblind, seeing the world in black and white, but she is nevertheless one of the world's leading neuroscientists, specializing in the neurophysiology of color vision. She knows everything there is to know about the neural processes involved in visual information-processing. But she does not know what it is like to see red. No amount of reasoning from the physical facts alone will give her this knowledge.

It follows that the facts about the subjective experience of color vision are not logically entailed by the physical facts. If they were, Mary could in principle come to know what it is like to see red on the basis of her knowledge of the physical facts. But she cannot. Perhaps Mary could come to know what it is like to see red by some indirect process: via an operation, perhaps, or through meditation, or by somehow exploiting her other senses. The point, however, is that knowledge of the facts about neurophysiology is not *sufficient*. Knowledge of all the physical facts will in principle allow Mary to derive all the facts about a systems reactions, and its various abilities and cognitive capacities, and even the content of a system's beliefs and desires; but she will still be entirely in the dark about its experience of red.

A related way to make this point is to consider systems quite different from ourselves, perhaps much simpler—such as bats or mice—and note that the physical facts about these systems do not tell us what their conscious experiences are like, if they have any (Nagel 1974 focuses on such considerations). Once all the physical facts about a mouse are in, the question of the nature of its conscious experience remains an *open question*: it is consistent with the physical facts about a mice that it has conscious experience, and it is consistent with the physical facts that it does not. From the physical facts about a bat, we can ascertain *all* the facts about a bat, except the facts about its conscious experiences. Knowing all the physical facts, we still do not know what it is like to be a bat.

Along similar lines we can consider a computer, designed as a simple cognitive agent (perhaps it has the intelligence of a dog), but that is similar to us in certain respects, such as its capacity for sensory discrimination. In particular it categorizes color stimuli in a manner quite similar to ours, grouping things that we would call "red" under one category and things we would call "green" under another. Even if we know every detail about the computer's circuits, it is an open question: (1) Is the computer experiencing anything at all when it looks at roses?; (2) If it is, it it experiencing the same sensory color quality that we have when we look at a rose, or some quite different quality? This is an entirely meaningful question, and knowing all the physical facts does not force one answer rather than another onto us. Therefore the physical facts do not logically entail the facts about conscious experience.

Argument 5: From the absence of analysis

A proponent of reductive explanation will no doubt assert that these arguments are all mere appeals to intuition. In a sense this is true. The arguments simply establish the *prima facie* plausibility and reasonableness of the claim that consciousness fails to logically supervene on the physical. Given that *prima facie* case, it is incumbent on the proponent of reductive explanation to give some idea of how the existence of consciousness might be entailed by the physical facts. While it is not fair to expect all the details, one at least needs an account of how such an entailment might *possibly* go.

Any attempt to demonstrate such an entailment is doomed to failure. For consciousness to be entailed by a set of physical facts, one would need some kind of analysis of the notion of consciousness—the kind of analysis whose satisfaction physical facts could imply. As far as I can tell, there is no such analysis.

The only analysis of consciousness that seems even remotely tenable for such purposes is a functional analysis. Upon such an analysis, it would be seen that all there is to the notion of something's being conscious is that it should play a certain functional role. For example, one might say that all there is to a state's being conscious is that it be verbally reportable, or that it be the result of certain kinds of perceptual discrimination, or whatever. However, these fail miserably as analyses. They simply miss what it *means* to be a conscious experience. Although conscious states may play various causal roles, they are not *defined* by their causal roles. Rather, what makes them conscious is that they have a certain phenomenal feel, and this feel is not something that can be functionally defined away.

To see how unsatisfactory these analyses are, note how they trivialize the problem of explaining consciousness. Suddenly, all we have to do to explain consciousness is explain our ability to make certain verbal reports, or to perform certain sorts of discrimination, or to manifest some other capacity. But on the face of it, it is entirely conceivable that one could explain all these things without explaining a thing about consciousness itself; that is, without explaining the *experience* that accompanies the report or the discrimination. To analyze consciousness in terms of some functional notion is either to change the subject or to define away the problem. One might as well define 'world peace' as 'a ham sandwich'. Achieving world peace will suddenly become much easier, but it will be a hollow achievement.

The alternatives to functional analysis look even worse. It is most unclear that there could be any other kind of analysis appropriate for reductive explanation. The only alternative might be a structural analysis – perhaps consciousness could be analyzed as some sort of biochemical structure—but that analysis would be even more blatantly false. Whether or not consciousness *is* a biochemical structure, that is not what 'consciousness' *means*. To analyze consciousness that way again trivializes the explanatory problem by changing the subject.

Certainly it is functional analyses that worked for the reductive explanation of biological phenomena, psychological phenomena, and so on. If there were to be a different kind of analysis, it would be something utterly new, and it is most unclear how its satisfaction could be implied by the facts about physics. All physics provides is a lot of causation and evolution among fields and particles in space-time. It is entirely unclear how all this causation could add up to something that logically entails the existence of consciousness. It seems that the notion of consciousness is more or less irreducible, being analyzable only in terms of concepts at the same level.

Note that this is quite unlike the sort of irreducibility that is sometimes supposed to hold for high-level concepts in general. We have seen that many high-level notions have no crisp definitions, and no manageable analyses in terms of necessary and sufficient conditions. Nevertheless, as we saw in the last chapter, these concepts at least have rough-and-ready analyses that get us into the ballpark, although they will inevitably fail to do justice to the details. Most importantly, it is straightforward to see that properties like life, learning, and so on can be analyzed as a functional property, even if spelling out the details of just *which* functional property is a difficult matter. Even though these properties lack crisp functional definitions, they are nevertheless quite compatible with entailment by the physical facts, as we have seen.

The problems with consciousness are in a different league. Here, the purported analyses do not even get into the ballpark. In a much starker way, they fail miserably to characterize what needs to be explained. There is no temptation to even *try* to add epicycles to a purported functional analysis of consciousness in order to make it satisfactory, as there is with similar analyses of life, learning, and other high-level notions. Consciousness is simply not to be characterized as a functional property in the first place. The same goes for analyses of consciousness as a structural property, or in other reductive terms. There is therefore no way for an entailment from physical facts to consciousness to get off the ground.

There are other arguments against logical supervenience that I will not go into here, some of which are more convincing than others. These include: (1) thought-experiments about the gradual disappearance of particular components of consciousness (Siewert 1993 provides a careful discussion of such cases); (2) consideration of phenomena such as blindsight—in such cases it appears that there are two hypotheses compatible with the physical facts: either that phenomenology is present but not reported, or that it is not present but unconscious discriminations are made; and (3) epistemological arguments about conditions under which we might be justified in believing that somebody had turned into a zombie (Kirk 1974 gives such an argument).

3.2 Objections

It seems to me that each of the five main arguments given above adequately demonstrates the failure of consciousness to logically supervene on the physical. If somebody finds one or even four of them unconvincing, then that is no problem, as it is good enough for my purposes if only one of them works. No doubt, however, some people will find that all five fail to establish their conclusion. In what follows, I will try to address some natural objections.

Objection 1: Couldn't one say the same thing about life?

A frequent response at this point is to argue that a vitalist might have used exactly the same argument. That is, a vitalist might have claimed that it is logically possible that a physical replica of me might not be *alive*, in order to establish that life cannot be reductively explained. (Dennett (1991) makes a point like this.) But the vitalist would have been *wrong*. By analogy, might not the advocate of zombies be equally wrong?

The cases are disanalogous in a straightforward way, however. All it *means* to be alive—if you like, all there is about life that requires explanation—is that a system have certain functional capacities, such as the ability to adapt, reproduce, and metabolize. Once one has explained the performance of these functions, one has explained all the phenomena that need to be explained. In the case of the mind, by contrast, there are certainly functional capacities that need to be explained—our various behavioral abilities, for instance—but there is also a further explanandum, namely conscious experience, which is precisely what is at issue.

Some unreconstructed vitalist might *claim* that there is something left out by a functional account of life, such as the vital spirit, but the obvious disanalogy here is that the vital spirit is not a manifest explanandum. In fact, there is no reason to believe that such a thing exists. Insofar as there is any reason to believe in vital spirit at all, it is as an explanatory construct—"we *must* have such a thing in order to be able to do such amazing stuff". As an explanatory construct, it can be eliminated when we find a better explanation. Conscious experience, by contrast, forces itself on one as an explanandum, and cannot simply be denied.

I think that this objection mischaracterizes the nature of the vitalist objection to physical explanations of life. The main consideration driving vitalist skepticism was a quite reasonable doubt about whether physical mechanisms were sufficient to perform all the *functions* associated with life. Given their ignorance of the vast capabilities of biochemical mechanisms, it is quite understandable that vitalists supposed that something extra, such as a vital spirit, might be required. It is notable that as physical explanation of the relevant functions gradually appeared, vitalist doubts mostly melted away. The vitalist skepticism is therefore entirely disanalogous to our worries about consciousness. The worries about consciousness do not concern the ability of physical mechanisms to perform various functions; rather, they concern the conceptual connection between the performance of those functions and consciousness.

Objection 2: But is *anything* logically supervenient on physics?

Another common response is to hold that a requirement of logical supervenience is too much to ask for reductive explanation, as all sorts of high-level phenomena are not logically derivable from microphysical facts. Perhaps no collection of facts about physics will entail the facts about architecture, biology, chemistry, demography, or economics.

I have addressed this issue at length in section 2.5. As we saw there, high-level facts in almost all such domains are logically supervenient on the physical. Most of the relevant properties are functional or structural properties, and such properties straightforwardly logically supervene.

It is interesting to note how most high-level facts—biological facts, for instance—evade the arguments that have been put forward about conscious experience. First, it is straightforwardly inconceivable that there could be a physical replica of a living creature that was not itself alive. At best, this might come about due to some context-dependence of aliveness, but fixing environmental facts would eliminate even that possibility. Second, there is no "inverted life" possibility analogous to the inverted spectrum. Third, when one knows all the physical facts about an organism (and possibly its environment), one has all the material one needs to know all the biological facts. Fourth, there is no epistemic asymmetry with life; facts about life in others are as accessible, in principle, as facts about life in ourselves. Fifth, the concept of life is plausibly analyzable in functional terms: to be alive is roughly to possess certain capacities to adapt, reproduce, and metabolize.

We can tell a similar story for almost any high-level phenomenon. Conscious experience is almost *sui generis* in its failure to logically supervene.

Objection 3: Isn't all this a bunch of circular intuitions?

There is certainly a sense in which all these arguments are based on intuition, but I have tried to make clear just how natural and plain these intuitions are, and how forced it is to deny them. The main intuition at work is that *there is something to be explained*—some phenomenon associated with first-person experience that presents a problem not presented by observation of cognition from the third-person point of view.

Given the premise that some explanandum is forced on us by first-person experience that is not forced on us by third-person observation, most of the arguments above fall out. It follows immediately, for instance, that what needs to be explained cannot be analyzed as the playing of some functional role, for the latter phenomenon is revealed to us by third-person observation, and is much more straightforward. Someone who argues that all we mean by 'consciousness' is the playing of some functional role is simply not talking about the same thing as I am.

Note that the premise here does not, at least initially, rule out the possibility that consciousness might "turn out" to be identical to some third-person-observable process. It merely asserts that there is an extra problem of explanation: a further *phenomenon* to be explained. This seems entirely obvious; it is the very *raison d'être* of the problem of consciousness. The only consistent way to get around these intuitions is to deny the phenomenon of conscious experience altogether. Some have taken this line (e.g., Dennett 1991) but it is a distinctly unpromising one. (I will directly address Dennett's eliminativist arguments elsewhere.)

Objection 4: What about new physics?

A fourth natural objection, and perhaps the most worrisome, is to note that in arguing that consciousness is not logically entailed by the physical facts about our world, we have been tacitly assuming that the physics of our world is something like physics as we understand it today: consisting in lots of particles and fields in the spatiotemporal manifold, undergoing complex processes of causation and evolution. An opponent might accept that nothing in *this* kind of physics entails the existence of consciousness, but argue that there might be a new kind of physical theory from which consciousness might fall out as a consequence.

It is difficult to evaluate this argument, as it is so vague and underspecified. One would at least like to see an example of how such a new physics might *possibly* go. Such an example need not be plausible in the light of current theories, but there would have to be a sense in which it would recognizably be physics. It seems to me that such a theory is most unlikely; and should it exist, it is unclear why we would call it a physical theory. If such a theory consists in a description of causation and evolution among fields, particles, and the like, then the usual problems will apply. And is it unclear what other kind of physical theory there could be.

Physics has certainly seen some radical developments, but these have preserved this basic theoretical framework. Relativity embedded the fields in a more interesting spatiotemporal geometry; quantum mechanics transmuted particles into waves and added a healthy dose of indeterminism, but nothing about these is close to the kind of physical revolution that would be required to entail consciousness. There was one truly radical development in quantum mechanics, centering around the role of the observer in measurement. This is still poorly understood. On some theories of measurement, consciousness itself plays a key role; however, such theories simply *assume* consciousness, and do not do anything to explain it.

Penrose (1989) suggests that a correct theory of quantum gravity might explain consciousness, but the suggestion is not spelled out. It is difficult to see why quantum gravity should be any more appropriate for this purpose than any traditional theory. Penrose appeals to the possibility that something non-algorithmic may be going on within quantum gravity. If this is so, then it may have consequences for theories of brain function, and for cognitive science in general, but it would be unlikely to make any difference in the explanation of consciousness.

Some have suggested that the *non-locality* of quantum mechanics, as suggested by recent experiments bearing on the Einstein-Podolsky-Rosen paradox and Bell's theorem, might be the key to a theory of consciousness (see Lahav and Shanks 1992 for suggestions along these lines). But even if these results do establish that physics is non-local, it is quite unclear why this should help in the explanation of consciousness. Given a non-local physical process, it seems equally coherent that the process could happen in the presence or the absence of consciousness. Consciousness still fails to logically supervene.

Of course, there must be some sense in which the physics of the universe must entail the existence of consciousness, if one *defines* physics as the fundamental science from whose facts and laws everything else follows. This construal of physics, however, trivializes the question involved. If one allows physics to include theories developed specifically to deal with the phenomenon of consciousness, unmotivated by more basic considerations, then we may get an "explanation" of consciousness, but it certainly will not be a reductive one. For our purposes, it is best to take physics to be the fundamental science developed to explain their observations about the external world.

If this kind of physics entailed the facts about consciousness, then consciousness would truly be reductively explained.

I suppose one has to allow the bare possibility that some new physics might be developed, motivated by external considerations, from which the existence of consciousness would fall out. In the absence of any notion of what such a physics would be like, however, and of how the entailment might possibly obtain, I think one can safely regard this possibility as implausible.

3.3 The failure of reductive explanation

The failure of consciousness to logically supervene on the physical implies that no reductive explanation of consciousness can succeed. Given any account of the physical facts purported to underlie consciousness, there will always be a further question: why are these facts accompanied by consciousness? For most other phenomena in the world, which logically supervene on the physical, such a question can be given a straightforward semantic answer—"because these facts imply that certain functions will be performed, and that is all it means to be alive" (or to learn, or whatever)—but no such answer will suffice for consciousness.

Rather, the fact that consciousness accompanies a given physical process is a *further fact*, not explainable simply by telling the story about the physical facts. In a sense, the accompaniment will have to be taken as a brute fact (although we can at least try to systematize the brute facts, as I will argue later). Although it is not impossible that one might get some kind of explanation in terms of a combination of (a) the underlying physical facts, and (b) the further facts that link the physical facts with consciousness, this explanation will no longer be a reductive one. The existence of consciousness will not be explained away, or deduced from more basic physical facts. Instead, it will be explained on its own terms.

Some might object that explanation of *any* high-level phenomena will postulate 'bridge laws' in addition to the details of a low-level theory, and it is using these bridge laws that the details of the high-level phenomena are derived. However, in such cases the bridge laws are not further facts about the world. As is carefully demonstrated by Horgan (1978), they are in principle logically supervenient on the low-level facts themselves (indeed they must be, if our account in Chapter 2 is correct). No such logical supervenience holds for consciousness, so there can be no correct reductive explanation. As Levine (1983) puts it, between physical theory and conscious experience there is an *explanatory gap* of a kind not found in other domains.

In what follows, I will illustrate this conclusion by considering various popular kinds of purported reductive explanations of consciousness, and explaining why they systematically fail.

3.4 Cognitive modeling

Cognitive modeling works well for most problems in cognitive science. By exhibiting a model of the causal dynamics involved in cognitive processes, and providing evidence that these dynamics match those in the case that needs to be explained, one provides an explanation of the causation of behavior in the subject under discussion (whether this is an individual, a species, or whatever). As we have seen, this provides a valuable kind of explanation for psychological phenomena, such as learning, memory, perception, control of action, attention, categorization, linguistic behavior, and so on. The reason for this is that if we have a model that captures the causal dynamics of someone who is learning, for example, it follows that anything instantiating those dynamics in the right environment will be learning. From the model we can see how certain functions are performed, and by the very meaning of 'learning', performance of those functions is all it takes to learn.

Such models, however, explain *only* the performance of functions. As we have seen, this is quite insufficient to explain consciousness. For any model we exhibit, it remains a further question why realization of the model should be accompanied by consciousness. This is not a question that description and analysis of the model alone can answer.

It is sometimes objected that purported models of consciousness are untestable, as there is no way to verify whether or not instantiations of the model are conscious. This is a problem, but there is a deeper problem. Even if we had (*per impossibile*) an "experience meter" that could peek in and tell us whether an instantiation was conscious, this would only establish a *correlation*. We would know that whenever the model is instantiated, consciousness goes along with it. But it would not *explain* consciousness, in the way that such models explain other mental phenomena.

Such models can certainly explain "consciousness" in the psychological senses thereof, where it is construed as some kind of cognitive or functional capacity. Many existing "models of consciousness" can be most charitably interpreted in this light. We can see these as providing explanations of reportability, or of attention, or of introspective abilities, and so on. None of them, however, gives us anything close to an explanation of why these processes should be accompanied by conscious experience.

As an illustration, I will briefly discuss some cognitive models that have been put forward as models of consciousness. For each of these, I will briefly present the model in question, and clarify the phenomena that the model really explains, or at least might explain. It is also interesting to note the varying attitudes of each of the authors to the key question of phenomenal experience.

1. *Baars*. Baars (1988) gives a book-length treatment of consciousness from the standpoint of cognitive psychology. He brings all sorts of experimental evidence to bear in establishing his main thesis: consciousness is a kind of *global workspace* in a

distributed system of intelligent information processors. When processors gain access to the global workspace, they broadcast a message to the entire system, as if they had written it on a blackboard. The contents of the global workspace are just the contents of consciousness.

Baars uses this model to explain a remarkable number of properties of human processing. Overall, this sort of model provides an excellent framework for explaining our access to information, and its role in attention, reportability, voluntary control, and even the development of a self-concept. The global workspace framework is therefore well-suited to explaining consciousness in its whole bundle of psychological senses; there is at least a general theory of *awareness* on offer here.

But there is no explanation of *experience* to be found here. The question of why writing of a message to some global workspace is sufficient to produce conscious experience is nowhere addressed in Baars' work. At best, his model explains certain functions associated with experience.

Baars (p. 27) addresses this sort of worry briefly: "A skeptical reader may...wonder whether we are truly describing conscious experience, or whether, instead, we can only deal with incidental phenomena associated with it." His response is to note that scientific theories tend to at least *approach* the "thing itself"; for instance, biology explains inheritance *itself*, and not just associated phenomena. But this is simply to ignore the ways in which consciousness is different in kind from these phenomena, as we have seen. With inheritance, various functions are all there is to explain. With consciousness, there is a further explanandum: experience itself. Baars' theory can therefore be seen as an interesting approach to the cognitive processes underlying consciousness, and one that gives us much indirect insight into consciousness, but it leaves the key questions—why is there consciousness and how does it arise from cognitive processing?—untouched.

2. *Dennett.* Dennett has put forward at least two distinct cognitive models of consciousness. The first of these (Dennett 1978c; see Figure 3) is a "box-and-lines" model, consisting in an account of the flow of information between various modules. Central to the model is (1) a perceptual module, (2) a short-term memory store M, which receives information from the perceptual module, (3) a control system, which interacts with the memory store by a question-and-answer-process, and which can direct attention to the contents of the perceptual module, and (4) a "public relations" unit, which receives speech act commands from the control system and converts them into public-language utterances.

What might this model explain? Although it is in a very simplified form (as Dennett would concede), it might be fleshed out to provide an explanation of *reportability*; that is, of our ability to report the contents of our internal states. It also provides the skeleton of an explanation of our ability to bring perceptual information to bear on

Figure 3: Dennett's cognitive model of consciousness (Dennett 1978, p. 155).

the control of behavior, to introspect our internal states, and so on. What it does not say anything about is conscious experience itself. The model tells us nothing about why all this processing should be accompanied by conscious experience.

Dennett (1991) puts forward a much more sophisticated account, appealing to much recent work in cognitive science. The model proposed there is essentially a "pandemonium" model, consisting in many small agents competing for attention, with the agent that "shouts the loudest" playing the primary role in the direction of later processing. On this model there is no central "headquarters" of control, but multiple channels exerting simultaneous influence. Dennett supplements this account with appeals to neuroscience, evolutionary biology, and connectionist models and production systems in artificial intelligence.

Despite the complexity of this account, it is directed at explaining more or less the same phenomena as the earlier account. If successful, it would provide an excellent explanation of reportability, and more generally of the influence of various sorts of information on the control of behavior. It also provides an interesting potential explanation of the focus of attention. This model goes no further than the previous model toward explaining conscious experience, however. While giving a provocative account of some of our *capacities*, it says nothing about why these capacities should be accompanied by *experience*.

Dennett addresses the question of conscious experience elsewhere, of course. Unlike most authors who put forward cognitive models, he explicitly argues that his models explain everything there is to explain, by arguing that the apparent phenomenon that is left out is a chimera. I will consider those arguments elsewhere.

3. *Johnson-Laird*. Johnson-Laird (1988) puts forward a computational account of consciousness, making an analogy between the processes that support consciousness and those found in digital computers.

> ...the computational architecture of the mind consists in a hierarchy of parallel processors; the processor at the top of the hierarchy is the source of conscious experience; this processor—the operating system— has access to a model of itself, and the ability to embed models within models recursively.

The self-modeling aspects of this model would seem to be especially appropriate as an explanation of *self-consciousness*, and of our ability to know about our internal states. The operating-system notion provides an interesting skeletal account of our control over our behavior, and of the way in which we have access to certain kinds of information in exerting that control. But nothing here says anything about why all this should be consciously experienced, as opposed to being accessible to later processes and exerting an effect on behavior. Indeed, Johnson-Laird ignores the question of phenomenal experience entirely. This should not be held against his account, which addresses various phenomena that have sometimes been called "consciousness", but the model cannot be seen as an explanation of consciousness in the really *difficult* sense.

4. *Shallice*. Shallice (1972) puts forward a model on which the brain contains a large number of *action systems*, only one of which can be maximally activated at a given time. Each action system has a certain kind of input, the *selector input*, which has two functions: it determines whether the action system will become dominant, and if so it sets the *goal* of the action system. Shallice identifies the content of consciousness with the selector input to the dominant action system.

This model is well-suited to explaining our selective access to information, and the way in which it is brought to bear in the control of behavior. The framework Shallice outlines is therefore quite promising as an explanation of what I called "awareness" in Chapter 1, or what Block (1991) calls "access consciousness". However, the problem of phenomenal experience is again untouched.

Shallice comes at phenomenal experience indirectly, arguing for his "identification" of selector inputs with consciousness by demonstrating certain isomorphisms in their structure. This isomorphism should not be surprising: as we saw in Chapter 1, there is a certain parallel between the contents of awareness and the contents of consciousness, so that one reflects the other (I will discuss this at much greater length in Chapter 6). However, merely noting this parallel does not in itself provide an *explanation* of consciousness. The question of why these processes of access give rise to experience remains as mysterious as ever.

Shallice puts forward some later models (e.g. Shallice 1988) that are more sophisticated, invoking for instance a "Supervisory System" and a language system, and making connections with contemporary psychological work on modules and schemas. The status of experience with respect to these models remains about the same, however. In this later discussion, Shallice talks only of a "correspondence" between consciousness and states in an information-processing model. This correspondence is quite reasonable and indeed a very useful notion, but nothing about the models *explains* the correspondence, or indeed explains why there is experience at all.

In general, these cognitive models can take us as far as the explanation of behavior, and can therefore indirectly explain various psychological properties—learning, memory, and the various psychological senses of "consciousness"—precisely because these properties are characterized by a role played in the causation of behavior. The models can even get us as far as a correspondence between certain psychological processes and conscious experience. What is not explained is *why* these psychological processes should be accompanied by conscious experience. The link between information-processing and the psychological aspects of mind is made clear, but the link between the psychological and the phenomenal aspects of mind – Jackendoff's (1987) "mind–mind problem" remains as problematic as ever.

3.5 Neurophysiological explanations

Neurophysiological approaches to consciousness have recently become popular, but they are just as problematic. At best, neuroscience can yield an account of the brain processes that are *correlated* with consciousness. Why these processes should give rise to consciousness is a further question, and one that neuroscience cannot answer. From the point of view of neuroscience, the correlation is simply a brute fact.

From a methodological standpoint, it is not obvious how one could even begin to come up with a neuroscientific theory. How would one perform the experiments that detect a correlation between some neural process and consciousness? What usually happens is that theorists implicitly rely on some psychological criterion for consciousness, such as the focus of attention, the control of behavior, and most frequently the ability to make verbal reports about an internal state. One then notes that some neurophysiological property is instantiated when these criteria are present, and one's theory of consciousness is off the ground.

The very fact that such indirect criteria are relied upon, however, makes it clear that no explanation of consciousness is on offer. At best, a neurophysiological account might be able to explain why the relevant psychological property is instantiated. The question of why the psychological property in question should be accompanied by conscious experience is left unanswered. Indeed, these theories rely on *assuming* a link between psychological properties and conscious experience, and therefore do nothing to explain that link. We can see this by examining some recent neuroscientific accounts of consciousness that have been put forward.

1. *Crick and Koch.* Much recent attention in neuroscience has focused on certain 40-hertz oscillations in the visual cortex and elsewhere. Crick and Koch (1990) have hypothesized that this sort of oscillations may be the fundamental neural feature responsible for conscious experience, and have advocated the development of a neurobiological theory along these lines.

Why 40-hertz oscillations? Primarily because evidence suggests that these these oscillations play an important role in the *binding* of various kinds of information into a unified whole. Two different kinds of information about a scene—the shape and location of an object, for instance—may be represented quite separately, but this theory suggests that the separate neural representations may have a common frequency and phase in their oscillations, allowing for the information to be bound together by later processes. In this way all sorts of disparate information might be integrated into the "contents of consciousness".

Such a theory might indeed provide a neurophysiological theory of *binding*, and perhaps eventually a theory of how information can be brought to bear in an integrated way in the control of behavior. But the key question remains unanswered: why should these oscillations be accompanied by conscious experience? The theory provides a partial answer: because these oscillations are responsible for binding. But the question of why binding itself should be accompanied by experience is not addressed. The theory gains the purchase it does by *assuming* a link between binding and consciousness, and therefore does nothing to explain it.

Crick and Koch seem sympathetic with the "big" problem of consciousness, calling it the "major puzzle confronting the neural view of the mind". They argue that pure cognitive-level approaches are doomed to be unsuccessful, and that neural-level theories are required. But they give us no reason to believe that their theory is better suited than the cognitive theories above when it comes to answering the really difficult questions. It suffers from just the same sort of problems. At best, the theory demonstrates a *correlation* between certain oscillations and experience; there is nothing in the way of an *explanation* of experience on offer.[4]

Figure 4: Edelman's scheme for primary and higher-order consciousness (Edelman 1992, p. 132).

2. *Edelman.* Edelman (1989) devotes an entire book to a neurophysiological theory of consciousness. His theory is too complex to summarize here, but it essentially involves re-entrant neural circuits by which perceptual signals can be conceptually categorized before they contribute to memory. Perceptual information and internal state interact in a subtle way (as diagrammed in Figure 4) to give rise to "primary consciousness". His model of "higher-order consciousness" brings in a new memory element through "semantic bootstrapping", which yields concepts of the self, past, and, future. All this is linked to language production through Broca's and Wernicke's areas.

Much of Edelman's book is devoted to the explanation of perception, memory, and language, rather than of consciousness. Insofar as it is devoted to consciousness, the

[4]In a later interview, Koch concedes that the theory may have little to say about the hardest questions. "Well, let's first forget about the real difficult aspects, like subjective feelings, for they may not have a scientific solution. The subjective state of play, of pain, of pleasure, of seeing blue, of smelling a rose—there seems to be a huge jump between the materialistic level, of explaining molecules and neurons, and the subjective level. Let's focus on things that are easier to study—like visual awareness. You're now talking to me, but you're not looking at me, you're looking at the cappuccino, and so you are aware of it. You can say, 'It's a cup and there's some liquid in it.' If I give it to you, you'll move your arm and you'll take it—you'll respond in a meaningful manner. That's what I call awareness." (Quoted in "What is Consciousness", *Discover*, November 1992, p. 96.)

discussion is often vague, but it seems that what ultimately might be explained by this sort of model is (1) perceptual awareness—that is, the effects of perceptual processing on later processes and on the control of behavior, and (2) self-consciousness, especially the origin of the concept of the self.

Edelman gives no account of how all this processing should give rise to conscious experience. He simply takes it that there is a correlation. He is up-front about this, noting that phenomenal experience is the hardest problem for a theory of consciousness, and that no physical theory will take us all the way to qualia.

> This suggests an approach to the problem of qualia. As a basis for a theory of consciousness, it is sensible to *assume* that, just as in ourselves, qualia exist in other conscious human beings, whether they are considered as scientific observers or as subjects. [...] We can then take human beings to be the best canonical referent for the study of consciousness. This is justified by the fact that human subjective reports (including those about qualia), actions, and brain structures and function *can all be correlated.* After building a theory based on the assumption that qualia exist in human beings, we can then look anew at some of the properties of qualia based on these correlations. It is our ability to report and correlate while individually experiencing qualia that opens up the possibility of a scientific investigation of consciousness. (Edelman 1992, p. 115.)

Edelman's theory therefore provides at best a *correlation* between brain processes and conscious experience. As before, it cannot *explain* the link between certain processes and experience, as it simply assumes such a link. Indeed, it is not clear that Edelman would even claim to be explaining phenomenal experience. Frequently, he simply claims to be explaining the processes that underlie conscious experience.[5] This is interesting and ambitious in its own right, but it does not yield a reductive explanation of experience itself.

3.6 Evolutionary explanations

Even people sympathetic with the above considerations are often drawn to the idea of an evolutionary explanation of consciousness. After all, consciousness is such a

[5]Edelman 1989, p. 168, is clear about this. "It is sufficient to provide a model that explains their discrimination, variation, and consequences. As scientists, we can have no concern with ontological mysteries concerned with *why* there is something and not nothing, or *why* warm feels warm." He makes an analogy with quantum field theory, which gives us a basis for discriminating energies and material states, but which does not tell us why there is matter in the first place. This analogy supports the non-reductive view, as we will see in the next chapter.

ubiquitous and central feature that it must have arisen during the evolutionary process for some *reason*; there must be some function that it serves. If we could get a clear enough idea of the relevant function, then we will have some idea why consciousness exists.

Such an idea, unfortunately, overestimates what an evolutionary explanation can provide us. The process of natural selection cannot distinguish between me and my zombie twin. Evolution selects properties according to their functional role, and my zombie twin performs all the functions that I perform just as well as I do; in particular he leaves around just as many copies of his genes. It follows that evolution alone cannot explain why conscious creatures rather than zombies evolved.

Some may be tempted to respond: "but a zombie *couldn't* do all the things that I can". But my zombie twin is by definition physically identical to me over its history, so it certainly produces indistinguishable behavior. Anyone wishing to question zombie capacity had best return to section 3.1, then, and find something wrong with the arguments there.

To see the point slightly differently, note that the real problem with consciousness is to explain the principles in virtue of which consciousness arises from physical systems. Presumably these principles—whether they are conceptual, metaphysical, or brutely nomic—are constant over space-time: if a physical replica of me had popped into existence a million years ago, it would have been just as conscious as I am. While evolution can be very useful to us in explaining why particular physical systems have evolved, it is utterly irrelevant to the explanation of the principles in virtue of which some of these systems are conscious.

3.7 Whither reductive explanation?

In my experience, many people agree with critiques of specific models like those above, but qualify this agreement: "Of course *that* model couldn't explain consciousness, but if we just wait a while an explanation will come along". I hope the discussion here has made it clear that the problems with this kind of explanation of consciousness are more fundamental than that. The problems with the models and theories presented here do not lie in the *details* (at least, we have not needed to consider the details in order to see what is wrong with them, although often separate criticisms could be mounted there). The problem lies in the overall explanatory strategy. These models and theories simply explain the wrong thing.

The problem with the reductive accounts is straightforward. Each of these accounts at best explains some cognitive *functions*. This is very useful for many purposes in cognitive science, but it is insufficient to explain consciousness. To explain function is not to explain consciousness.

Optimists about reductive explanation often make an analogy with life, holding that a few centuries ago we would have had no idea what a reductive explanation of life might look like, either. But the analogy does not hold. Explaining life is only a matter of explaining certain functions, and that was so even then. It was simply that a few centuries ago, nobody had any good explanations of the functions in question. As we have seen, the problem with reductive explanation of consciousness is much more serious. Even when we have a good explanation of the relevant functions, it will still not constitute an explanation of consciousness. It is inevitable that increasingly sophisticated reductive "explanations" of consciousness will be put forward, but these will only produce increasingly sophisticated explanations of cognitive functions. Even such "revolutionary" developments as the invocation of connectionist networks, nonlinear dynamics, and artificial life will ultimately only provide more powerful explanations of various functions. This may make for some very interesting cognitive science, but the mystery of consciousness will not be removed.

It might be supposed that there could eventually be some reductive explanatory technique that explained something other than function, but it is very hard to see how this could be possible. Such an explanation would be radically different in kind from those outlined above. In any case, the arguments in 3.1 rule such a possibility out. Given that consciousness fails to logically supervene, the existence of consciousness will always be a further fact relative to any explanation appealing only to physical facts, and so will be unexplained by such an account.

For an explanation of consciousness, then, we must look elsewhere. We certainly need not give up on *explanation*; we need only give up on an over-optimistic hope for a *reductive* explanation. I will outline the form that a non-reductive explanation of consciousness might take toward the end of the next chapter, and I will begin the task of fleshing out such an explanation in the remainder of this work.

Chapter 4

An Argument for Dualism

4.1 The falsity of physicalism

For the most part, I have been concerned with the explanation of consciousness and not with its ontology. Nevertheless, it is probably clear by now that the arguments in the last chapter have significant ontological consequences. In particular, they directly imply a version of dualism. The argument for this proceeds as follows.

(1) In our world, there are conscious experiences.

(2) There is a logically possible world physically identical to ours, but in which the facts about conscious experience in our world do not hold.

(3) Therefore facts about conscious experience are further facts about our world, over and above the physical facts.

(4) Therefore physicalism is false.

As in 2.1, I take physicalism to be the doctrine that the physical facts about the world exhaust all the facts, in that every particular fact is entailed by the physical facts.

Given the possibility of a physically identical zombie world, it follows that the presence of consciousness is an "extra" fact about our world, not guaranteed by the physical facts alone. The contingency and arbitrariness of our world is not exhausted by the contingency of the physical facts; there is extra contingency due to the presence and nature of consciousness. To use a phrase due to Lewis (1990) and Seager (1991), consciousness carries phenomenal *information*. The physical facts are incomplete as a constraint on the way the world is; the facts about consciousness constrain it further.

The argument works equally well if we appeal to the possibility of a world with *inverted* conscious experiences. Such a world is physically identical to ours, but not all the facts about conscious experience in our world hold. It follows that the facts about conscious experience in our world are further facts over and above the physical facts, so that physicalism is false.

Either way, if consciousness does not logically supervene on the physical, then physicalism is false. The failure of logical supervenience implies that some fact about our world does not hold in a physically identical world, so that it is a further fact over and above the physical facts. Using the Kripke's image from Chapter 2: when God created the world, after ensuring that the physical facts held, *he had more work to do.* He had to ensure that the facts about consciousness held. The possibility of physically identical zombie worlds or inverted worlds shows us that he had a choice. The world might have lacked experience or it might have contained different experiences, compatibly with the physical facts. The fact that experiences are as they are is therefore independent of the physical facts.

What kind of dualism?

This argument does not force us into a Cartesian substance dualism, positing a separate realm of mental substance that exerts its own influence on the physical. The best evidence of contemporary science tells us that the physical world is more or less causally closed: for every physical event, there is a physical sufficient cause. If so, there is no room for a mental "ghost in the machine" to do any extra causal work. A small loophole is opened up by the existence of quantum indeterminacy, but this seems insufficient for a non-physical mind to do significant causal work, and in any case such a "mind of the gaps" seems quite implausible (I will discuss this issue further in section 4.5). Given that the physical world is near enough to being causally closed, then in principle behavior should be fully explainable in physical terms.

The dualism implied here is instead a kind of *property* dualism: conscious experience in an individual is a property that is not entailed by the physical facts about that individual. At the moment, that is all we can say about it: to declare it a substance or even a "thing" would be to go beyond what is justified by the argument. All we know is that there are properties of individuals and perhaps of mental states—the phenomenal properties—that are ontologically independent of physical properties.

To see this another way, note that every property can be seen as an intension, picking out those individuals possessing the property in a given possible world. A physical property, even a high-level physical property, will pick out corresponding individuals in microphysically identical worlds. But we have seen that a phenomenal property such as 'has conscious experience' picks out a non-empty reference class in our world, but picks out an empty reference class in a microphysically identical zombie world. So phenomenal properties are not physical properties of any kind.

Although consciousness may not be logically *entailed* by the physical facts, it may well *arise* from a physical substrate. Indeed, it is very plausible that in the actual world, any replica of me would have indistinguishable conscious experiences. It follows that there is a lawful relation between the physical and consciousness. Consciousness supervenes on the physical facts; it is just that the supervenience relation is not a logical relationship. Instead, it is a contingent, nomic relationship. Consciousness arises from the physical facts in virtue of certain contingent laws of nature, which are not themselves implied by physical laws. This position is best summed up by calling it *nomic supervenience* of consciousness on the physical.

This position is implicitly held by many people who think of themselves as physicalists. It is common to hear "of course I'm a materialist; the mind certainly arises from the brain". The very presence of the word "arises" should be a tip-off here. One tends not to say "learning arises from the brain", for instance—and if one did, it would be in a temporal sense of "arises". Rather, one would more naturally say that learning *is* a process in the brain. The difference in phrasing is significant. Edelman (1992) similarly subtitles his (purportedly non-dualist) book: "How the Mind Originates in the Brain".

There is little point arguing at length about just what qualifies as physicalism, but it seems to me that the existence of further contingent facts over and above the physical facts is a significant enough modification to the received physicalist world view to deserve a different name. Certainly, if all that is required for physicalism is that all facts be lawfully connected to the physical facts, then physicalism becomes a weak doctrine indeed.

On the other hand, those who hold this kind of dualism may be temperamentally much closer to physicalists than to Cartesian dualists in many ways, due to their belief in the physical causation of behavior. Conversely, most of the bad press that dualism has received is due to the stronger interactionist varieties. One hears frequently that the successes of cognitive science and neuroscience make dualism an implausible view; but these successes only count toward the likeliness of physical explanations of behavior, and so do not distinguish between the physicalist and the nomic supervenience views.

The argument against the logical supervenience of consciousness has been given in the previous chapter, and objections were dealt with then. I will now deal with objections to the argument from the failure of logical supervenience to the falsity of physicalism.

4.2 Objections from *a posteriori* necessity[1]

A common response upon hearing this argument is to object that we have only established that a zombie world is *logically* possible, which is quite different from being "metaphysically possible", or possible *tout court.* Whereas conceptual coherence suffices for logical possibility, metaphysical possibility is more constrained. This objection is usually accompanied by an appeal to Kripke's *Naming and Necessity* (1980), which demonstrates the existence of necessary truths such as 'water is H_2O' whose

[1]This section is philosophically technical. Non-philosophers and those not especially interested in *a posteriori* necessity can safely skip to the next section.

necessity is only knowable *a posteriori*. In the sense of these objectors, 'water is H_2O' is "logically possible" but not "metaphysically possible".

This appeal to *a posteriori* necessity turns out to be something of a red herring. We can see this best by using the framework for dealing with *a posteriori* necessity developed in 2.4.[2]

Recall that on the Kripkean framework there are two intensions (functions from possible world to reference) associated with any concept: a prior intension which fixes reference in the actual world, and a posterior intension which picks out reference in counterfactual worlds. The prior intension associated with 'water' is something like 'watery stuff'. The posterior intension is 'H_2O', which is derived from the prior intension by applying Kaplan's *dthat* operator: '*dthat*(watery stuff)' is equivalent to 'H_2O' in all possible worlds, as watery stuff is H_2O in the actual world.

'Logical possibility' comes down to the possible truth of a statement when evaluated according to the prior intensions involved. The prior intensions of 'water' and 'H_2O' differ, so it is logically possible in this sense that water is not H_2O. 'Metaphysical possibility' comes down to the possible truth of a statement when evaluated according to the posterior intensions involved. The posterior intensions of 'water' and 'H_2O' are the same, so it is metaphysically necessary that water is H_2O.

The objection to our argument, then, comes down to the objection that in using arguments from conceivability and the like, we have demonstrated the possibility of a zombie world using the *prior* intensions of the notions involved, but not using the more appropriate *posterior* intensions. While the prior intension of phenomenal notions may not correspond to that of any physical notion, the posterior intensions may be the same. If so, then phenomenal and physical/functional concepts may 'pick out' the same properties *a posteriori*, despite the *a priori* distinction.[3]

The easiest way to see that none of this affects the argument for dualism is to note that the argument goes through if we use the prior intension throughout, and ignore the posterior intension. We saw in Chapter 2 that it is the prior intension that is most relevant to explanation, but it also serves us well in the argument for dualism. For note that whether or not the prior and posterior intensions coincide, the prior intension determines a perfectly good property of objects in possible worlds.[4] 'Watery stuff' is a perfectly reasonable property, even though it picks out a different class from 'H_2O' in some possible worlds. If we can show that there are possible worlds that are physically identical to ours but in which that property is lacking, then the argument

[2]Related arguments for the irrelevance of *a posteriori* necessity to these considerations can be found in Jackson 1993, Lewis 1994, and White 1986.

[3]By this analysis, we can see that objections to my argument based on an appeal to 'psychofunctionalism' (see Block 1980b) and other theories about the *a posteriori* equivalence of phenomenal and physical/functional properties fall into this class. These objections are equally vulnerable to the remarks below.

for dualism will be complete.

This is just what we have done with consciousness. We have demonstrated that there are worlds just like ours that lack consciousness, according to the prior intension thereof. This difference in worlds is sufficient to show that there are properties of our world over and above the physical properties. By analogy: if we could show that there were worlds physically identical to ours in which there was no watery stuff, we would have established dualism about water just as well as if we had established that there were worlds physically identical to ours in which there was no H_2O. And importantly, the difference with respect to the prior intension can be established independently of *a posteriori* factors, so that considerations about *a posteriori* necessity are irrelevant.

The point can be made even more compelling when we note that with consciousness, the prior and posterior intensions coincide. What it takes for a state to be a conscious experience in the actual world is for it to have a phenomenal feel; what it takes for something to be a conscious experience in a counterfactual world is for it to have a phenomenal feel. There is no *dthat* in the concept of consciousness corresponding to the *dthat* in the concept of water; or if there is, it is entirely redundant. The *dthat* in the concept of water reflects the fact that there could be something that looks and feels like water in some counterfactual world that in fact was not water, but merely watery stuff. But if something feels like a conscious experience, even in some counterfactual world, it *is* a conscious experience. All it means to be a conscious experience, in any possible world, is to have a certain feel. (Kripke himself makes a similar point, although he puts the point in terms of essential properties rather than in terms of meaning.)

Even if someone implausibly insists that there is a *dthat* affecting the notion of consciousness, the argument will still go through. We simply focus on the prior intension used to fix reference, as above. For instance, if 'consciousness' comes to '*dthat*(has a phenomenal feel)', then we simply focus on the intension 'has a phenomenal feel'. The arguments in Chapter 3 establish that there is a possible world in which my replica lacks a phenomenal feel, so the property of having a phenomenal feel is a fact over and above the physical facts, and the argument for dualism is successful.[5]

The most general way to make the point is to note that nothing about Kripke's *a posteriori* necessity renders any logically possible worlds impossible. It simply points

[4]Strictly speaking the prior intension determines an *index-relative* property of an object in a possible world (or a relation between objects and indices), as the prior intension applies to *centered* possible worlds that come supplied with such an "index" location. This index-relativity cannot be exploited to help our objector, however, as all the arguments go through even when the index location is included in the supervenience base. For instance, even if Mary's facts about the world include facts about where *she* is located, this will not help her know about red any better.

[5]Jackson (1980) makes a similar point, arguing that even if *a posteriori* considerations can establish the physicality of the property *pain*, the problem still arises from the property *pain-presents*.

out that some of them are misdescribed, because we are applying terms according to their prior intensions rather than the more appropriate posterior intensions. One might have thought it logically possible *a priori* that water is XYZ, rather than H_2O. In conceiving this, one imagines something like a world in which XYZ is the clear liquid found in oceans and lakes. However, Kripke's analysis shows us that due to the way the actual world turns out, we are misdescribing this world as one in which XYZ is water, as we are describing it with the prior intension instead of the more appropriate posterior intension. Instead, it is one in which XYZ is watery stuff. Such considerations cannot show the impossibility of this apparently possible world; they simply show us the correct way to describe it.[6]

As we saw in 2.4, Kripkean *a posteriori* considerations show us that the posterior intension $F_a : W \rightarrow R$ differs from the prior intension $f : W \rightarrow R$. This puts some *a posteriori* constraints on the application-conditions of concepts, but the relevant space of worlds stays constant throughout. There are two kinds of possibility of *statements*, but there is only one relevant kind of possibility of *worlds*.

Someone who wanted to use considerations about *a posteriori* necessity to defeat the argument about dualism would therefore have to show us that we are misdescribing the world we are conceiving of as the 'zombie world'. I have argued above that the prior and posterior intensions of 'consciousness' are identical, so this could not be the case, but in any case it does not matter as long as this world is lacking *something* that our world has, whether or not we call it 'consciousness'. And this is precisely what the arguments in Chapter 3 establish.

A final tack that a physicalist might take here is to argue that in claiming that the world in question is *physically identical* to ours, we are misdescribing it. Just as the XYZ-world seems to contain water but does not, the world in question might *seem* physically identical while being physically different. This might seem forced, but I can see one way it could go. An opponent might argue that there are certain properties essential to the physical constitution of the world that are not accessible to physical investigation. In conceiving of a "physically identical" world, we are really only conceiving a world that is identical from the standpoint of physical investigation, while differing in the essential inaccessible properties, which are also the properties that guarantee consciousness.

For something to qualify as an electron, on this view, it might not be sufficient that it be causally related to other physical entities in the way that an electron is. Perhaps some hidden essence of electronhood is also required. And perhaps the presence of consciousness in our world is guaranteed by these hidden facts. To do

[6]It follows that there is nothing especially metaphysical about 'metaphysical necessity'. It is merely a brand of conceptual necessity with an *a posteriori* semantic twist, arising from the two-dimensional nature of our concepts. For more on the theme that *a posteriori* necessity reflects convention as much as metaphysics, see Putnam (1983) and Sidelle (1989; 1992).

the job, the essence in question would have to be very different from more familiar physical properties; but perhaps the essence is somehow proto-phenomenal in nature. The world that we are conceiving might lack these hidden essential properties, and therefore might fail to be physically identical to ours. It would satisfy the prior intensions of physical predicates, which presumably apply on the basis of superficial extrinsic properties, but not the posterior intensions, which require a certain hidden essence.

It seems to me that this argument is based on a misuse of physical predicates, which apply even *a posteriori* on the basis of extrinsic relations between physical entities, irrespective of any "hidden" properties. This is a purely conceptual question: if electrons in our world have hidden proto-phenomenal properties, should we call an otherwise identical counterfactual entity that lacks those properties an electron? It seems to me that we should. But even if we allow that such hidden properties can be conceptually constitutive of physical properties, the difference between this point of view and the property dualism I have been advocating would be merely semantic. It would remain the case that the world has properties that are not fixed by the properties that physics reveals to us. After ensuring that a world is identical to ours from the standpoint of our physical theories, God has to expend further effort in making that world identical to ours *tout court*. The dualism of "physical" and "non-physical" properties would be replaced on this view by a dualism of "accessible" and "hidden" physical properties, but nothing important would change.

Brute metaphysical necessity

We have established that the Kripkean considerations about reference and necessity do not have any force against the argument; if the zombie world is logically possible, these considerations cannot count against its metaphysical possibility. Some may still assert, however, that the world might be metaphysically impossible *nevertheless*. On such a view, there is a modality of metaphysical impossibility that is distinct and more constrained than logical possibility, and one that arises for reasons independent of the Kripkean considerations. Thus there are fewer metaphysically possible worlds than there are logically possible worlds, and the metaphysical impossibility of statements can arise for reasons independent of the semantics of the terms involved.

On this view, there are worlds that are quite conceivable, even according to the strongest strictures on conceivability, but which are not possible at all. Certainly there are various kind of gaps between conceivability and possibility elsewhere, as with Kripkean arguments about the conceivability of 'water is XYZ', and indeed with arguments about the conceivability of 'π is rational'; but these, as we have seen, do not rule out the possibility of any conceivable *world*. They are merely instances where we misdescribe such worlds by some statement. On the current position, 'zombie world' does not misdescribe the world that we are conceiving, even according to a posterior intension; it is just that that world is not metaphysically possible.[7]

The short answer to this objection is to argue that there is no reason to suppose that such a modality should exist. Such "metaphysical necessities" will put constraints on the space of possible worlds that are utterly brute and inexplicable. It may be reasonable to countenance brute, inexplicable facts about *our* world, but the existence of such facts about the space of possible worlds would be quite bizarre. Assertion of such facts would be a matter of arbitrary stipulation. One might as well stipulate that it is metaphysically impossible that stones should move upward when one lets go of them.

Someone who holds that a zombie world is logically possible but metaphysically impossible has to answer the key question: *Why couldn't God have created a zombie world?* Presumably it is in God's powers, when creating the world, to do anything that is logically possible. Yet the advocate of metaphysical necessity must say either: (1) the possibility is coherent, but God could not have created it, or (2) God could have created it, but it is nevertheless metaphysically impossible. The first is quite unjustified, and the second is entirely arbitrary. If the second holds, in any case, our argument against physicalism will still be strong enough to go through, as it will follow that after fixing the physical facts about the world, God still had more work to do.

The only reason to postulate such metaphysical constraints would be to make physicalism come out true. It would violate all the usual principles by which we determine whether something (a scientific law, say) is necessary or contingent. If somebody insists on taking this line, I suppose I cannot stop them, but the view ends up looking not much different from dualism. Instead of having nomic relations linking consciousness and the physical, one has brute "metaphysical" relations, relations which seem contingent but are in fact necessary. Such a view has the virtue of saving physicalism, perhaps, but at the cost of making it utterly mysterious how consciousness *could* be physical.

The much more natural view is that there are only two objective grades of necessity and possibility of worlds: logical necessity and nomic necessity. Postulating a third intermediate variety is quite arbitrary.

4.3 Further objections

Objection 1: Nomic supervenience and physicalism.

---

[7]Few have explicitly taken this position in print, although it seems to me that Bigelow and Pargetter (1990), Levine (1993), and Loar (1990) are implicitly committed to a position like this. Horgan has advocated the position in personal communication.

As I said earlier, the nomic supervenience view is held implicitly by various people who think of themselves as physicalists. Some people may hold that this view is better described as 'physicalism' than as dualism: after all, it allows that consciousness *arises* from the physical. While little rests upon this terminological issue, I have already given reasons why such a view is much more naturally described as dualism. Nomic supervenience implies that there are facts over and above the physical facts, contingent relative to those facts, and this is sufficient to imply the falsity of physicalism on any reasonable construal thereof.

Searle (1992) has advocated a view similar to mine in the relevant respects, but describes his view as a variety of physicalism. Like me, Searle suggests that microphysical states and consciousness are linked only by a nomic connection. He allows that a zombie replica is logically possible, holding only that consciousness is *caused* by states of the brain. He denies that the view is a variety of dualism, arguing that a similar story holds elsewhere. For instance, H_2O causes liquidity, and no-one is a dualist about liquidity.

This is a false analogy, however. Given all the microphysical facts about a particular batch of H_2O, it is logically impossible that those facts could be satisfied without liquidity being instantiated. The notion of a non-liquid replica of a batch of liquid H_2O is simply incoherent. All it means to be liquid is to have certain structural and dispositional properties, and instantiation of these properties is logically entailed by the relevant microphysical facts. It follows that the relation between the microphysical facts and liquidity is not causal but logical. The microphysical facts do not *cause* liquidity; they *constitute* it. So there is a relevant disanalogy between consciousness and liquidity, and Searle's position fails.

Objection 2: Supervenience and identity

It might be objected that the claim that a zombie world is possible simply begs the question against physicalism. If conscious states are identical with physical states, as physicalism claims, then it is simply impossible to have the same physical states without having the same conscious states.

The case for the logical possibility of zombies has been made on its own merits, however, independent of the question of whether conscious states and physical states are identical. Of course it is true that if zombies are logically possible, then there is a clear sense in which conscious states and physical states cannot be identical,[8] but this no more begs the question than any argument begs its conclusion. An objection to the argument must find something wrong with the arguments for logical possibility. One cannot simply assert identity as a brute fact; that would come to the brute "metaphysical necessity" view criticized above.

[8]There is perhaps a sense in which my view is compatible with an identity between phenomenal

I have deliberately avoided talk of identity in running the argument, running it in terms of supervenience instead. This is partly because I find the justification-conditions for claims of identity obscure, but it is also because in these matters supervenience is a more fundamental issue than identity. Claims of supervenience can be evaluated without bringing identity into the picture, and if supervenience fails then identity fails.[9]

The use of supervenience rather than identity, incidentally, is what qualifies this argument as something other than a mere rerun of Descartes'. Descartes argued, roughly: I can conceive that my mind exists without my body, or that my body exists without my mind; therefore my mind cannot be identical to my body. This argument is flawed in various ways. In general, conceivability of non-identity does not imply non-identity (witness heat and molecules), unless certain strong restrictions are placed. Furthermore, it is not clear that materialism requires any identity claims; it is sufficient that the mental be constituted by the physical (see Boyd 1980), or that it supervene appropriately on the physical (see Haugeland 1982). The argument against supervenience is quite distinct from the argument against identity, although it is reminiscent in various ways, and it avoids the problems with Descartes' argument.

Objection 3: Two aspects of the same thing?

A common view of consciousness is that conscious states and brain states are the same states, but under two different modes of presentation. Physical states are what we see when we look at the brain from the outside, and conscious states are what we see when we look from the inside. The underlying states might be the same, just as the Morning Star and Evening Star are both Venus, presented by two different modes. On this view, then, consciousness might be entirely physical.

The reply to this objection is a straightforward question: why does the internal aspect exist, and why is it the way it is? My zombie twin is presumably quite capable of introspecting his internal state, but for him this introspection does not feel like anything at all. The double-aspect theorist must either tell us why the existence and nature of the conscious aspect follow logically from the physical facts, in which case we are back to the arguments in Chapter 3, or must concede that the conscious aspect is contingent relative to those facts, in which case the view collapses into a variety of dualism.

state-tokens and physical state-tokens, as long as the state-tokens in question can have distinct physical and phenomenal properties. At least, views like mine have sometimes been understood this way. This seems awkward to me, but in any case I do not have strong feelings about the matter one way or the other; a description of my view in terms of supervenience is clearer than any description in terms of identity.

[9] In general, a relevant slogan is *no identification without explanation*, although unpacking this takes some work.

There may well be some sense in which consciousness and the physical are two aspects of the same thing. Indeed, I will advocate such a view later. However, it is inescapable that given only the facts about the physical aspect, the existence of the conscious aspect is not fully determined. This view is therefore best described as a variety of dualism rather than as physicalism.

Objection 4: Emergence.

It is often held that consciousness is an "emergent" phenomenon, emerging as a result of many low-level physical interactions. On some construals of emergence, the high-level phenomena are not deducible from the low-level facts, but this is nevertheless held to be compatible with physicalism. Common examples include the emergence of self-organization in biological systems, or the emergence of flocking patterns from simple rules in simulated birds (see for example Langton 1990, or Reynolds 1987).

As far as I can tell, there are actually two varieties of emergence in common parlance. In the first, logical supervenience fails, but so does physicalism. The second is compatible with physicalism, but logical supervenience holds. The first variety is the more traditional philosophical notion of emergence, advocated by Alexander (1920), among others. On this view, when certain low-level elements are combined, high-level phenomena (such as consciousness) come into existence in a way that is not predictable from the low-level events at all, even in principle. This view is a variety of dualism, for the reasons I have given already.

The second variety has become popular more recently in artificial intelligence, artificial life, and other areas. In these areas, it often happens that striking, unexpected high-level phenomena result from the interaction of low-level processes following simple rules. Such phenomena are said to be 'emergent'. Unexpected though these phenomena are, however, they nevertheless follow logically from the low-level facts. It may not be straightforward to derive them from the low-level *laws* that are set up, but they follow in principle, and more importantly, they follow straightforwardly from the collection of low-level *facts*. Indeed, such emergent phenomena are frequently observed as the output of computational processes. Such outputs are certainly logical consequences of the processes involved, even if an explicit derivation is messy.

Almost every example of emergence that we see in the natural world—the emergence of self-organization, for instance—is emergent in the second sense. But consciousness is different. If someone wishes to argue that consciousness is emergent in the second sense, they need to provide an in-principle argument showing why it is logically determined by the low-level facts, and again we are back to Chapter 3. If consciousness is emergent at all, it is emergent in the first sense; but this collapses into a variety of dualism rather than physicalism.

4.4 Relation to other arguments for dualism[10]

The argument from logical possibility is of course not original with me.[11] Indeed, I think it is the fundamental anti-materialist argument in the philosophy of mind. Nevertheless, it has not received the careful attention it deserves. More attention has focused on two anti-materialist arguments by Kripke (1972) and Jackson (1982). These arguments strike me as related, but as perhaps less fundamental. Jackson's argument is important for the entry it provides to the argument from logical supervenience, and what is right about Kripke's argument is derivative on what is right about the argument I have presented, as we will see.

Jackson's argument

I have already discussed Jackson's argument, the Knowledge Argument, in the context of establishing the failure of logical supervenience, where it plays a supporting role. Recall that the argument is concerned with Mary, a colorblind neuroscientist, who knows all the physical facts about color processing in the brain. Later, when she gains color vision, she learns some new facts—namely, she learns what it is like to see red. The argument concludes that the physical facts do not exhaust all the facts, and that physicalism is false.

This argument is closely related to the argument from zombies, in that both revolve around the failure of phenomenal facts to be entailed by physical facts. In a way, they are flip sides of the same argument. As a direct argument against materialism, however, Jackson's argument is often seen as vulnerable due to its use of the intensional notion of knowledge. Many attacks on the argument have centered on this intensionality—arguing, for instance, that the same fact can be known in two different ways. These attacks fail, I think, but the best way to see this is to proceed directly to the failure of supervenience, which is cast in terms of metaphysics rather than epistemology. Nevertheless, I think that Jackson's argument is a good argument. The framework developed here helps bring out just why the various objections to it fail. I will briefly discuss some of these objections.

First, various respondents have argued that although Mary gains new knowledge upon seeing red, this knowledge does not correspond to any new *fact*. She simply comes to know an old fact in a new way, because of the intensionality of knowledge

[10] This section is philosophically technical and can be skipped.

[11] The most explicit version of the argument from zombies is given by Kirk (1974). It is also present in Campbell (1970), Nagel (1974), Robinson (1976), and many other places. My presentation of the argument differs mostly in (1) the use of the notion of supervenience to provide a unifying framework, and (2) consideration of the role of *a posteriori* necessity. See also Seager (1991) for a related argument from the possibility of inverted spectra.

(Churchland 1985; Horgan 1984b; McMullen 1985; Tye 1986). For example, Tye appeals to the intensional difference between 'this liquid is water' and 'this liquid is H_2O': in a sense these express the same fact, but one can be known without the other. Similarly, Churchland appeals to the gap between knowledge of temperature and knowledge of mean kinetic energy, Horgan discusses the difference between knowledge of Clark Kent and knowledge of Superman, while McMullen points to Mark Twain and Samuel Clemens.

These gaps arise precisely because of the difference between prior and posterior intensions. One can know things about water without knowing things about H_2O because the prior intensions differ—there is no *a priori* connection between water-thoughts and H_2O-thoughts. Nevertheless, in a sense there is only one set of facts about the two: because of the *a posteriori* identity between water and H_2O, the relevant posterior intensions coincide.[12] In the terminology used earlier, 'If this is water, it is H_2O' is logically contingent but metaphysically necessary. This objection therefore comes to precisely the same thing as the objection from the distinction between logical necessity and (Kripkean) metaphysical necessity discussed earlier, and the discussion there of prior and posterior intensions is sufficient to refute it.

To see the point a different way, note that someone who knows the whole physical story about H_2O will know all about water. The H_2O-facts will include or imply the fact that H_2O is found in lakes and oceans, is a clear liquid, is found in our environment (though see the third objection below), and so on. In short, the full physical story will include all the information required to see that H_2O is water. The same goes for the full physical story about heat and temperature, or about Superman and Clark Kent. In these cases, therefore, physical information is exhaustive. But even knowing the entire physical story about neurophysiology will not enable one to know all about conscious experience, so the analogy fails.

A second, more sophisticated objection (Loar 1990, and perhaps Bigelow and Pargetter 1990) also holds that Mary gains new knowledge of old facts because of intensionality, while allowing that Kripkean considerations alone cannot demonstrate this. In our terminology Loar recognizes that the objection above cannot do the job for the materialist, as it allows a distinction in prior intension between physical and phenomenal notions; the anti-materialist can simply rerun the argument with the property corresponding to the prior intension, as I did earlier. As Loar puts it: even though 'heat' and some statistical-mechanical predicate *designate* the same property (posterior intension), they nevertheless *introduce* distinct properties (prior intension).

[12] There is a sense in which water-facts can differ from H2)-facts; a sense in which "water is wet" and "H_2O is wet" express different facts, as for that matter do "water is H_2O" and "H_2O is H_2O". In this sense, we individuate facts by the prior intensions of the terms used to express them. It is more common to individuate facts by the posterior intensions of the relevant terms, however; at least, that is the sense in which water-facts and H_2O-facts are the same facts.

So he takes the argument further, and argues that two predicates can introduce the same property—that is, share the same prior intension—even when this sameness is not knowable *a priori*. If so, then Mary's knowledge of phenomenal properties may just be knowledge of physical/functional properties, even though she could not have connected the two beforehand.

How can two prior intensions coincide without our being able to know it *a priori*? Only if the space of possible worlds is smaller than we would have thought *a priori*. We think the intensions differ because we conceive of a world where they have different reference, such as a zombie world. Loar's position therefore requires this world is not really possible, despite the fact that we cannot rule it out on conceptual grounds, and despite the fact that Kripkean considerations cannot do any work for us. This position therefore comes to precisely the same thing as the second "metaphysical necessity" objection considered above. Like that objection, Loar's position requires that a conditional from physical facts to phenomenal facts be metaphysically necessary despite being logically contingent, where this gap cannot be explained by a difference in prior intensions. Like that objection, Loar's position requires a brute and arbitrary restriction on possible worlds. Loar offers no argument for this restriction, and his position is subject to precisely the same criticisms.

A third objection makes an analogy between Mary's plight and indexical knowledge (McMullen 1985; Bigelow and Pargetter 1990). Although Mary gains new knowledge, it is argued that this is no more puzzling than other cases where someone who knows all the relevant objective facts discovers something new: e.g. an omniscient amnesiac who discovers '*I* am Rudolf Lingens', or a well-informed insomniac who does not know that it is 3:49 a.m. *now* (see Perry 1977 and Lewis 1979). But this is not a useful analogy. As we saw in Chapter 2, indexicals accompany facts about conscious experience in their failure to logically supervene on physical facts, but they are all settled by the addition of a thin "indexical fact", about the spatiotemporal location of the agent in question. By contrast, we can allow Mary perfect knowledge about her indexical relation to the rest of the world, and her knowledge of red experiences will not be improved. In lacking phenomenal knowledge, she lacks far more than someone lacking indexical knowledge.

A fourth objection argues that Mary gains no new knowledge upon seeing red, but merely gains an ability, such as the ability to imagine, or to recognize, or to predict (Lewis 1990; Nemirow 1990). On the face of it, this is manifestly implausible—no doubt Mary does gain some abilities, as when she learns to ride a bicycle, but it certainly seems that she learns something else: some *facts* about the nature of experience. For all she knew before, experience might have been like this, or like that, or even like nothing at all; now she knows that it is like *this*. No such knowledge comes along when one learns to ride a bicycle (except, of course, for knowledge about the phenomenology of bicycle-riding). Still, this is probably the only reasonable answer that a materialist can give: we have seen that materialist replies that concede that Mary gains factual knowledge all fail. This position is at least consistent, but it suffers from all the usual difficulties with functional analyses of phenomenal concepts.

Kripke's argument

I will discuss Kripke's argument in some detail in what follows, going through what I consider to be wrong and right about it, and arguing that the parts that are right correspond directly to this argument; whereas those components of Kripke's argument that do not correspond to anything in this argument are the parts that fail.

Kripke's argument has been perhaps the most influential argument against materialism, although it is actually directed at a particular form of materialism, namely the contingent identity thesis developed by Place (1957)and Smart (1959). It can, however, be seen to have a broad force against all forms of materialism. The argument runs approximately as follows.

According to the contingent identity thesis, certain mental states (such as pains) and brain states (such as C-fibers firing) are identical, even though 'pain' and 'C-fibers firing' do not *mean* the same thing. The identity here is supposed to be contingent, rather than necessary, just as the identity between water and H_2O is contingent. Against this, Kripke argues that *all* identities are necessary: if X is Y, then X is *necessarily* Y (as long as the terms X and Y designate "rigidly", picking out the same individual or kind across worlds). Water is *necessarily* H_2O, he argues; that is, water is H_2O in every possible world. The identity may *seem* contingent—that is, it might seem that there is a possible world in which water is not H_2O but XYZ—but this is illusory. In fact, the possible world that one is imagining contains no water at all. It is just a world in which there is some *watery stuff*—stuff that looks and behaves like water—made out of XYZ. In asserting that this watery stuff is water, one is misdescribing it.

Similarly, Kripke argues, if pains are identical to the firing of C-fibers, then this identity must be necessary. But the identity certainly does not *seem* to be necessary. On the face of it, one can imagine a possible world where a pain occurs without any brain state whatsoever (disembodied pain), and one can imagine a world in which C-fibers fire without any accompanying pain (in a zombie). Further, he argues, these possibilities cannot be explained away as merely apparent possibilities, in the way that the possibility of water without H_2O was explained away. For that to be the case, we would have to be *misdescribing* the "disembodied pain" world as one in which pain occurred, when really there was just "painy stuff" (i.e., stuff that feels like pain) going on. Similarly, we would have to be misdescribing the zombie as lacking pain, when all it really lacks is painy stuff. On such an account, the zombie would presumably have real pain, which is the firing of C-fibers; it is just that it doesn't feel like real pain.

But this cannot be the case, says Kripke: *all it is* for something to be pain is for it to feel like pain. There is no distinction between pain and painy stuff, in the way that there is a distinction between water and watery stuff. One could have something that felt like water without it being water, but one could not have something that felt like pain without it being pain. Pain's feel is *essential* to it. Therefore these possible worlds really are possible, mental states are not necessarily identical to brain states, and therefore cannot be identical to brain states at all.

In fact Kripke runs the argument in two different ways: once against token-identity theories, and once against type-identity theories. Token-identity theories hold that *particular* pains (such as my pain now) are identical to particular brain states (such as the C-fibers firing in my head now). Kripke argues in the above fashion that a particular pain could occur without the particular associated brain state, and vice versa, so they cannot be identical. Type identity theories hold that mental states and brain states are identical as *types*: pain, for example, might be identical as a type to the firing of C-fibers. Kripke holds that this is straightforwardly refuted by the fact that one could instantiate the mental state-type without the brain state-type, and vice versa. Overall, we can count four separate arguments here, if we split them according to the target (token- or type- identity theories) and according to the method of argument (from the possibility of disembodiment or from the possibility of zombies).

There are some obvious differences between Kripke's argument and the argument I have given. For a start, Kripke's argument is couched entirely in terms of identity, whereas I have avoided that notion, relying instead on the notion of supervenience. Secondly, Kripke's argument is closely tied to his theoretical apparatus involving rigid designators and *a posteriori* necessity, whereas that apparatus plays only a secondary role in my argument, in answering certain objections. Thirdly, Kripke's argument is usually seen to rely on a certain essentialism about various states, whereas no such doctrine is invoked in my argument. Fourthly, my argument is not about "mental states" in general, but only about phenomenal states; this is only a minor point, however, as I doubt that Kripke intended his argument to apply to non-phenomenal mental states such as beliefs. Fifthly and importantly, my argument nowhere appeals to the possibility of disembodiment, as Kripke's does.

Nevertheless there are obvious similarities. Both are modal arguments, involving necessity and possibility in key roles. And both appeal to the logical possibility of zombies. I will now go through what succeeds and fails in Kripke's arguments, arguing that the part that succeeds corresponds directly to the argument I have given.

The arguments against token-identity are generally held to be inconclusive. This is partly because they rely on intuitions about what counts as *that very thing* across possible worlds, and such intuitions are notoriously unreliable. Kripke's claim that one could have *that very* pain-state without *that very* brain-state relies on the claim that what is *essential* to that pain-state is its feel, and only its feel. But such claims about the essential properties of individuals are notoriously unclear. The token-identity theorist can respond by arguing that it is just as plausible that the firing of C-fibers is an essential property of the state. Of course, C-fiber firing does not seem to be essential to pain as a *type*, but who is to say that it is not essential to this particular pain-token, especially if that token is identical to a brain state? If it is, then one simply could not have the particular pain in question without the particular brain-state. A line like this is taken by Feldman (1974), who argues that painfulness need not be essential to a particular pain, and by McGinn (1977), who argues that both painfulness *and* C-fiber firing might be essential to a particular pain. If so, then in imagining a disembodied version of my pain, one is not imagining *that very* pain but merely a duplicate of it. The same goes for imagining my C-fiber firing without pain. So the arguments against token-identity are inconclusive, although the arguments against type-identity survive.

Next, the argument from disembodiment does not establish a conclusive case against materialism. It is true that it may refute a type-identity thesis of the kind put forward by Place and Smart, but materialism does not require such a thesis. As Boyd (1980) notes, the materialist need not hold that mental states are physical states in all possible worlds—it is compatible with materialism that in some worlds mental states are constituted out of non-physical stuff, as long as in *this* world they are physically constituted. The possibility of disembodiment only establishes the possibility of dualism, rather than its truth. To illustrate this, we can note that that few would argue that the possibility of non-physical life implies dualism about biology.

This leaves the argument from the possibility of zombies, which curiously is the part of Kripke's argument that has received the least critical attention, with most commentators focusing on the possibility of disembodiment.[13] Of course, the argument that zombies yield against strong type-identity theses may be irrelevant, due to the fact that materialism does not require such a thesis, but there is a more general argument lurking here. The possibility of zombies, Kripke argues (p. 153-4) shows that even after God created all the physical stuff going on when one has a pain, such as a brain with C-fibers firing, *he had to do more work* in order that those firings be felt as pain. This is enough to establish that physicalism is false.

This argument is almost exactly the argument I have given against physicalism, and I think that it is the only conclusive part of Kripke's argument. Unfortunately it has generally been overlooked amidst all the discussion of identity theses, disembodiment, and the like.[14] Even Kripke assigns this phrasing of the argument a non-central

[13]Kripke does not use the term 'zombie', but his argument comes to the same thing.

role.

To summarize, it seems to me that insofar as Kripke's argument is correct: (1) the possibility of disembodiment is inessential; (2) phrasing the argument in terms of identity is inessential and in fact detrimental; (3) essentialism is inessential, except insofar as the feel of pain is essential to pain as a type—but that is just to say that what 'pain' *means* is 'something that feels like pain'; (4) Kripke's apparatus of rigid designation and the like is not central, although it is required to answer a particular kind of objection.[15] However, his discussion contains a sound (if overlooked) argument against materialism in the argument from the logical possibility of zombies, via the argument that "God has to do more work".

4.5 Is this epiphenomenalism?

A problem with the view I have advocated is that at least on the face of things, if consciousness is merely nomically supervenient on the physical, then it lacks causal efficacy. The physical world is more or less causally closed, in that for any given physical event, it seems that there is a physical explanation (modulo a small amount of quantum indeterminacy). This means that there is no room for a non-physical consciousness to do any independent causal work. It seems to be merely an epiphenomenon, hanging off the engine of physical causation, but making no difference in the physical world. It exists, but as far as the physical world is concerned it might as well not. Huxley (1874) advocated such a view, but many people find it counterintuitive and repugnant. Indeed, this consequence is enough to cause Kirk (1979) and Seager (1991) to question the conclusions of their arguments, and to consider the possibility that consciousness might logically supervene on the physical after all.

This argument can be formalized, and such a formalization is given by Kirk (1979),

[14] In his careful analysis, Boyd (1980) notes that the possibility of zombies, unlike the possibility of disembodiment, entails the falsity of materialism. He therefore provides a separate argument against this possibility, but the argument is sketchy and unconvincing. Boyd makes an analogy with a computer computing a particular function, arguing (1) that it may *seem* to us that one could have all the circuits of the computer just as they are without that function being computed, but that this is nevertheless impossible, and (2) that the apparent possibility of zombies is analogous to this. However, the analogy fails. The situation with the computer is analogous to the (tenuous) "apparent possibility" that there might be a physical replica of me that does not learn what I learn, or does not perceive what I perceive. Nothing in this analogy can account for the far more compelling nature of the apparent possibility of a replica without conscious experience.

[15] Although the argument is often taken to be an application of Kripke's theory of rigid designation, a version of it could in principle have been run ten years earlier, before the theory was developed. One could have asked Place and Smart why the physical facts about H_2O necessitate that it be water (or watery), whereas the physical facts about brain states do not seem to necessitate that they be painful.

Horgan (1987), and Seager (1991). If we assume (1) that the physical world is causally closed, and (2) that consciousness causes some physical events, then it follows under certain natural assumptions about causation that consciousness must logically supervene on the physical.[16] Therefore the mere nomic supervenience of consciousness implies that consciousness is epiphenomenal, given that the physical world is causally closed.

There are a number of things one can say about this. Firstly, finding a conclusion counterintuitive or even "repugnant" is not sufficient reason for rejecting that conclusion. Epiphenomenalism may be counterintuitive, but it is not *obviously* false. So if a sound argument forces it on us, we should accept it, although it may provide motivation to go back and look more closely at the argument. If failure of logical supervenience implies epiphenomenalism, then this may make logical supervenience *desirable*, but desirability does not imply truth. One cannot simply assert logical supervenience as a brute fact. Logical supervenience requires some account of how the physical facts might entail the facts about consciousness, and this, I have argued, cannot be done.

Secondly, some might argue that this view need not imply epiphenomenalism, if one accepts a certain kind of Humean (regularity-based) account of causation. For instance, nomic supervenience is perfectly compatible with the claim that pain sensations, *ceteris paribus*, tend to be followed by withdrawal reaction. On some Humean views, this is sufficient for it to be the case that pain causes withdrawal reactions. I find a Humean account of causation implausible for the reasons given in 2.5, so I will not investigate the option further.

Still, note that this at least gives us an idea of why consciousness *appears* to play a causal role. There are all sorts of systematic regularities between conscious experiences and later physical events, of the sort that lead us to suppose a causal connection. Indeed, we can note in a Humean fashion that the *evidence* for a causal connection is never greater than that for a corresponding law-governed regularity. A large part of the grounding for our intuition that conscious experience is causally efficacious can therefore be explained away in such a manner.

Third, this view does not put all aspects of pain out of the loop. It remains the case that *psychological* pain plays a definite causal role, in virtue of its logical supervenience on the physical (indeed, this aspect of pain is characterized precisely by its causal role). It is merely the phenomenal aspects of pain that seem to be epiphenomenal. Nomic supervenience is compatible with the view that every mental state has two aspects, a psychological and a phenomenal aspect, where it is the psychological aspect that

[16] Horgan reaches a slightly different conclusion: that consciousness must metaphysically supervene on the physical. But this involves the dubious invocation of a metaphysical necessity that is not grounded in conceptual necessity, with the consequent existence of brute, inexplicable constraints on the space of possible worlds, as we saw in section 4.2.

does the causal work. There might even be a sense in which conscious experience has a kind of *derivative* causal efficacy, derivative on the psychological state that supports it. It merely lacks *independent* causal efficacy. Nevertheless, it may still seem unsatisfactory that experience can only be causally efficacious in this derivative sense. One might hope that the phenomenal aspects of pain, say, have causal efficacy in their own right.

Some people, persuaded by the arguments for dualism given earlier, but convinced that phenomenal consciousness must play a causal role, may be tempted by an interactionist variety of dualism. I think that this would be a mistake. It would be placing a hefty bet on the future of science, one that does not currently seem at all plausible; physical events seem inexorably to be explained in terms of other physical events. It would be placing a large wager on the future of cognitive science in particular, implying that the usual kinds of physical/functional models will be insufficient to explain behavior.

Before moving on, however, I should make a couple of points about causal closure and quantum mechanics. The first concerns the existence of quantum indeterminacy. This implies that the physical world is not fully causally closed, in that physical events need not be determined by preceding physical events. This has caused some to suppose that a non-physical consciousness might do its causal work by filling these causal gaps, determining which value some physical magnitudes might take within an apparently "probabilistic" distribution. Although these decisions would have only a tiny proximate effect, it might be argued that they could eventually have large effects on behavior, due to "chaotic" effects whereby tiny fluctuations are amplified to make a difference at the macroscopic scale. (Eccles 1986 presents a theory along these lines.)

Although this is an audacious and interesting suggestion, that there are a number of reasons why it is unlikely to work. Firstly, the theory contradicts the quantum-mechanical postulate that these microscopic "decisions" are entirely random, and in principle it implies that there should be some detectable pattern to these—a testable hypothesis. Secondly, in order that this theory allow that consciousness do any *interesting* causal work, it needs to be the case that the behavior produced by *these* microscopic decisions is somehow different in kind than that produced by most other sets of decisions that might have been made by a purely random process—presumably the behavior is more rational than it would have been otherwise, or it leads to remarks such as "I am seeing red now" that the random processes would not have produced. This again is testable in principle, by running a simulation of a brain with real random processes determining those decisions. Cognitive science does not provide any reason to suppose that such processes would lead to qualitatively different behavior. Thirdly, in order for the mind to be able to make microscopic decisions that lead to the right kind of behavior, it would have to solve, implicitly or explicitly, an almost intractable backward-chaining problem: how do we get from the desired behavior to the right microscopic fluctuation needed to produce it? Perhaps an omnipotent god could do this, but it would seem an unlikely thing to be built into the capacities of a mind.

A second way in which quantum mechanics bears on the issue of causal closure lies with the fact that on some interpretations of quantum theory, consciousness itself plays a vital causal role—it is required to perform the so-called "collapse of the wave-function". This collapse is supposed to occur upon any act of measurement; and on one interpretation, the only way to distinguish a measurement from a non-measurement is via the presence of consciousness. This theory is certainly not universally accepted (for a start, it *presupposes* that consciousness is not itself physical, surely contrary to the views of most physicists), and I find it quite implausible myself, but in any case it seems clear that the kind of causal work consciousness performs here is quite different from the kind required for consciousness to play a role in directing behavior (although see Lahav and Shanks 1992 for a contrary view). For a start, the "collapse" is supposed to occur in *external* objects that are perceived, not within the brain itself. This theory is generally silent on what is happening within the brain. Secondly, even if consciousness does somehow manage to "collapse" the brain state (by introspection, perhaps?), then all the above remarks about apparently random processes and their connection with behavior will still apply.

The points about quantum mechanics, and interactionist dualism in general, are rendered academic by the observation that there seem in principle to be relatively straightforward physical/functional explanations of most of our *behavior* related to consciousness, including the things we say about consciousness, and such explanations need not invoke the existence of consciousness at all. (Of course, the existence of such an explanation of our behavior, and of the things we say, does not imply an explanation of consciousness itself.) This point will be developed at length in later chapters.

All this being said, however, I do not think that nomic supervenience necessarily forces us into an epiphenomenalistic view. At a first pass it seems to, but there may be more subtle developments that enable us to reconcile nomic supervenience with out intuitions about causal efficacy. One alternative might be a kind of dual-aspect theory, wherein physical stuff and conscious experience are seen to be two different aspects of some broader kind. If so, then it is really the broader kind that enters into causation at the fundamental level. Physical stuff and consciousness will both be causally efficacious, in a manner derivative on their status as aspects of the broader kind. One such possibility will be presented in future work (summarized in the appendix), where I will sketch a theory on which the physical and the phenomenal are seen to be different aspects of information, and on which information is fundamental.

The most promising way to reconcile failure of logical supervenience with causal efficacy, however, may be to note that as we saw in Chapter 2, there are precisely two broad classes of facts that do not logically supervene on particular physical facts: facts about consciousness, and facts about causation itself. It is very natural to speculate that these two failures might be intimately related, and that consciousness and causation have some deep metaphysical tie. Both consciousness and causation are quite mysterious, after all, and two mysteries might be more neatly wrapped into one. Perhaps consciousness itself is some kind of causal nexus; perhaps it is constitutive of Hume's "unknowable causal relation". Perhaps the relationship is more subtle. On such a view, consciousness will not have the kind of causal efficacy that a billiard ball has, but who expected that it would? Instead, its intimate relation to causation itself might give it efficacy of a more subtle kind. I will go into this matter in more depth in later work. For now, we can simply note that the mere failure of logical supervenience does not make the causal efficacy of consciousness entirely hopeless.

I will therefore not describe my view as epiphenomenalism, although I recognize that it may turn out to imply epiphenomenalism. Causation itself is simply too ill-understood to make such a judgment. It may be that a final theory of consciousness will require significant advances in our understanding of causation. The label "epiphenomenalism" also tends to suggest a picture on which consciousness largely floats free of cognition, whereas I would like to advocate a picture on which they are intimately related (very loosely, instead of thinking of consciousness floating up in the sky, think of it sitting down in the causal cracks)

Even if my view turns out to be a version of epiphenomenalism, I think the arguments for nomic supervenience are sufficiently compelling that one should accept them. Epiphenomenalism is counterintuitive; but the alternatives are not just counterintuitive, they are *wrong*, as we have already seen, and will see again in the next section. The overall moral is: If the arguments suggest that nomic supervenience is true, then we should learn to live with nomic supervenience.

Some may not find this satisfactory, holding that the epiphenomenalistic nature of this view is a fatal flaw. I have some sympathy with this position—what it amounts to is the implication that when it comes to consciousness, *all* the alternatives are bad. If such a person comes away with the feeling that consciousness is simply an utter mystery, then that is not completely unreasonable. However, I do not think that the problems with nomic supervenience make the alternatives any better; all of them have fatal problems. The problems with nomic supervenience are less damaging—the view is counterintuitive, but the other views are just *wrong*. In the next section, I will describe what I see as the possible alternative positions on the ontological status of consciousness, and sum up what is wrong with each of them.

4.6 The logical geography of the issues

The argument for my view, in summary, is an inference from roughly four premises:

(1) Conscious experience exists.
(2) Conscious experience does not logically supervene on the physical.
(3) If there are facts that do not logically supervene on the physical facts, then physicalism is false.
(4) The physical world is causally closed.

(1), (2), and (3) clearly imply the falsity of physicalism. This, taken in conjunction with (4) and the plausible assumption that physically identical beings will have identical conscious experiences, implies the view that I have called nomic supervenience: conscious experience arises from the physical according to some laws of nature, but is not itself physical. The various alternative positions can be catalogued according to whether they deny (1), (2), (3), or (4). Of course some of these premises can be denied in more than one way.

Denying (1):

(i) Eliminativism. On this view, there are no facts about conscious experience. Nobody is conscious in the phenomenal sense. (The view allows for psychological consciousness).

Denying (2):

Premise (2) can be denied in various ways, depending on how the entailment in question proceeds—that is, depending on which physical properties are centrally responsible for entailing consciousness. I call all of these views 'reductive physicalist' views, because they suppose an analysis of the notion of consciousness that is compatible with reductive explanation.

(ii) Reductive behaviorism. On this view, what it means to be conscious is to be disposed to behave in certain ways. Consciousness is therefore entailed by the behavioral properties of an organism. It follows that a zombie replica of me is logically impossible, as a replica would share all my behavioral dispositions.

(iii) Reductive functionalism. This view takes consciousness to be entailed by physical states in virtue of their functional properties, or their causal roles. On this view, what it means for a state to be conscious is for it to play a certain causal role. In a world physically identical to ours, all the relevant causal roles would be satisfied, and therefore the conscious states would all be the same. The zombie world is therefore logically impossible.

(iv) Non-functionalist reductive physicalism. On this view, the facts about consciousness are entailed by some physical facts in virtue of their satisfaction of some non-functional property. Possible candidates might include biochemical and quantum properties.

(v) Holding out for new physics. According to this view, we have no current idea of how physical facts could explain consciousness, but that is because our current conception of physical facts is too narrow. When one argues that a zombie world is logically possible, one is really arguing that all the fields and particles interacting in the void, postulated by current physics, could exist in the absence of consciousness. However, a new physics, invoking a radically different theoretical framework, might be sufficient to explain consciousness.

Denying (3):

(vi) Brute physicalism. This is the view that although there may be no logical entailment from the physical facts to the facts about consciousness, and therefore no reductive explanation of consciousness, consciousness *just is* physical: the physical facts "metaphysically necessitate" the facts about consciousness. So the zombie world is impossible, even though it is logically possible.

(vii) Nomic physicalism. This view holds (as does mine) that consciousness is connected to the physical facts only by nomic necessity, and not by any stronger form of necessity—that is, consciousness *arises* from the physical—but claims that this is compatible with physicalism.

Denying (4):

(viii) Interactionist dualism. This view accepts that consciousness is non-physical, but denies that the physical world is causally closed, so that consciousness can play an autonomous causal role.

Then there is my view, which accepts (1), (2), (3), and (4):

(ix) Nomic supervenience. Consciousness nomically supervenes on the physical, without logically (or "metaphysically") supervening.

There is also a tenth common view, which is generally underspecified:

(x) Don't-have-a-clue physicalism: "I don't have a clue about consciousness. It seems utterly mysterious to me. But it must be physical, as physicalism must be true." Such a view is held widely, but rarely in print (although see Fodor 1992).

Of these views: the first seems to be manifestly false, requiring a proponent to be in the grip of a philosophical ideology; (ii), (iii), and (iv) rely on false analyses of the notion of consciousness, and therefore change the subject; (v) and (viii) may stand a chance of being right, but they place large empirical bets on the way that physics will turn out, bets that seem quite implausible; (vi) either makes an invalid appeal to Kripkean semantics, or relies on a bizarre metaphysics; and (vii) is straightforwardly false. I have a certain amount of sympathy with (x), having a strong temperamental inclination toward physicalism, but presumably it must eventually reduce to some more specific view, and none of these seem to work.

This leaves (ix), which seems to be the only view without a fatal flaw: indeed, it is entailed by four well-justified premises. Its main problem is that it may imply a kind of epiphenomenalism. However, even if the view has this consequence, epiphenomenalism is merely counterintuitive, and this counterintuitiveness does not rule out the view as definitively as the others are ruled out; and in any case, I will argue later that the view need not be as epiphenomenalistic as it may seem.

Inevitably, some people will find the dualistic nature of my position unpalatable, and will be searching for a way out. I will argue soon that dualism is not nearly as unreasonable as one might have thought, but for now I will briefly resurvey the best materialist options and summarize why none of them succeed. (I also survey some views that do not explicitly fall into this framework in an appendix to this chapter, spelling out how they fit the framework and what their problems are.)

Of the materialist positions above, I think the only *remotely* tenable options are (i), (iii), (v), and (vi). The first, eliminativism, can be ruled out as manifestly at odds with the evidence. Admittedly the only argument here is an appeal to the reality of experience, but this is quite sufficient. As we have seen, conscious experience is not an environmental phenomenon that can be directly measured; it is an internal phenomenon whose existence has to be verified by one person at a time, independently. But the internal evidence is quite enough, at least in my case, for me to know that there is something it is like to be me, and that I really do have red sensations. I presume it is the same with you. Eliminativism is such a manifestly implausible position that one can only mount a substantial argument against it when one is given specific arguments *for* it in the first place. I will address some such arguments, put forward by Daniel Dennett, in a later chapter.

Alternative (v), holding out for new physics, has a certain appeal to it. I certainly cannot refute it directly, in the absence of any conception of what that new physics might look like. I will content myself with the observation that such physics would have to look *vastly* different from current physical theories, moving beyond patterns of causation and evolution among various particles, fields, and state-spaces to something else entirely. It is currently beyond our conception how such a theory could simultaneously be a physical theory and entail the existence of consciousness; and current physics gives no reason to suppose that things will move in such a direction.

This option has to be left on the table, but I think it must be considered a long shot.

Alternative (vi) holds that even though there is no logical supervenience, phenomenal properties are still physical properties, because facts about conscious experience are tied to physical facts by brute "metaphysical necessity". We have seen that this position cannot be supported by an appeal to Kripke's account of *a posteriori* necessity. Instead, it requires brute constraints on the space of possible worlds. There is no prior support for these constraints, which are simply invoked to make a desired conclusion fall out (materialism, or perhaps the causal efficacy of consciousness). The best response to this, I think, is simply to argue that possibility doesn't work that way. If a world is internally consistent, then it is possible.

Even if this position is accepted, it ends up looking remarkably similar to mine. It remains the case on this position that there is no physical *explanation* of conscious experience; and it remains the case that conscious experience is connected to the physical by a contingent-seeming supervenience connection. It is just that this account calls that connection 'metaphysical' rather than 'nomic'. Certainly, everything I say later in developing a theory of the connection between consciousness and the physical will apply equally to this position. The differences between the positions are mostly cosmetic.

This leaves alternative (iii), the analytic functionalism of Armstrong and Lewis, which I think is ultimately the only serious materialist contender. Leaving aside various wild options: if materialism is true, then consciousness logically supervenes, and the only halfway reasonable way for it to logically supervene is via a functional analysis. On this view, then, all it *means* for something to be a conscious experience is for it to play a certain causal role in a system. Phenomenal properties are treated exactly the same way as psychological properties, such as learning or categorization.

The fundamental problem with this view is that it misrepresents what it means to be a conscious experience, or to be conscious. When I talk about my red sensations, or about my aural qualia, or about the phenomenology of emotion, I am not talking about any ability, or any functional capacity. I am simply talking about the way the state feels. When I wonder whether some other being is conscious, I am not wondering about their abilities, which I may know all about already; I am wondering whether there is something it is like to be them. This view ultimately rests on a misanalysis of the concepts involved, and provides support for materialism only by changing the subject.

There are various indirect ways to see this. One way is to note, as we did earlier, how this approach trivializes the problem of explaining consciousness. All we need to do to explain consciousness, on this view, is to explain certain functional capacities. But it seems entirely coherent that one could explain these capacities without explaining consciousness. Once the capacities are explained, there is a further explanandum, namely consciousness.

A second way is to note that this analysis collapses the distinct notions, outlined in Chapter 1, of awareness and consciousness. Presumably if consciousness is to be functionally analyzed, it will be analyzed roughly as we analyzed awareness then: in terms of a certain accessibility of information to later processing, which can be rationally utilized in the control of behavior. Awareness is certainly a valid concept, but it is quite distinct from the concept of conscious experience; it is a far more straightforward thing. The functionalist treatment collapses the two notions of consciousness and awareness into one, and therefore does not do justice to our conceptual system.

A third way is to note that with any functional analysis of a phenomenon, there will be a certain degree of semantic indeterminacy about just what counts as an instance of that phenomenon. Does a mouse have beliefs? Can bacteria learn? Is a computer virus alive? The best answer to these questions is usually: in a sense yes, in a sense no. It all depends on how we draw the boundaries in the concepts of 'belief', 'learning', and 'life'; there are no canonical boundaries to draw. There will always be some degree of vagueness in any such high-level functional concept, and any boundaries will to some extent be a matter of stipulation.

With consciousness, however, there is no room for such indeterminacy. Does a mouse have conscious experience? Does a virus? These are not matters for stipulation. With conscious experience, there is a fact of the matter: either there is something it is like to be a mouse, or there is not. It is not up to us to define the mouse's experience into or out of existence. (Note, by contrast, that the question of whether a mouse or a virus is *aware* is very much a matter for stipulation.) To be sure, there is probably a continuum of conscious experience from the very faint to the very rich; but if something has conscious experience, however faint, we cannot stipulate it away. Any functional analysis that was capable of reflecting this determinacy would have to be remarkably precise, unlike the functional analyses found elsewhere in psychology and science. No such precise functional analysis could be derived from our concepts alone. It follows that the notion of consciousness is not something that can be functionally analyzed.

I do not doubt that some people will retain analytic functionalism, because the cost of giving it up, dualism, is too high a price to pay. This is understandable, even if it is putting the ideological cart before the conceptual horse. But it is not a view that is capable of taking consciousness seriously.

Reflections on dualism

Many people, including a past self of mine, have thought that they could simultaneously take consciousness seriously and remain a materialist. In this chapter I have argued that this is not possible, and for straightforward reasons. Any materialist account of consciousness must either deny the phenomenon (through eliminativism), or change the subject (through functionalism). The moral: those who want to come to grips with the phenomena must embrace a form of dualism. One might say: You can't have your materialist cake and eat your conscious experience too.

Nevertheless, at this point many will be casting about for an alternative to the position I have put forward, because they find its dualistic nature unacceptable. This reaction is natural, given the various negative associations of dualism, but I think it is not grounded in anything more solid than contemporary dogma. To see this, it is worthwhile considering the various *reasons* that one might have for preferring materialism to dualism, and measuring the force of these reasons in the current dialectical situation.

The first reason to prefer materialism is *simplicity*. This is a good reason. Other things being equal, one should prefer a simpler theory over one that is ontologically profligate. Ockham's razor tells us that we should not multiply entities without necessity. But other things are not equal, and in this case there *is* necessity. We have seen that materialism cannot account for the phenomena that need to be explained. Just as Maxwell sacrificed a simple mechanistic worldview by postulating electromagnetic fields in order to explain certain natural phenomena, we need to sacrifice a simple physicalistic worldview in order to explain consciousness. We have paid due respect to Ockham's razor by recognizing that for materialism to be overthrown, one will need good arguments; but when the arguments against materialism are there, Ockham cannot save it.

The second and perhaps the most pervasive reason to believe in materialism is inductive: materialism has always worked elsewhere. We have materialist accounts of life, cognition, and the weather; or where we lack such accounts, we have reason to suppose that they are not far off. Why should consciousness be any different?

But this reason is easy to defeat. As we have seen, there is a simple explanation for the success of materialist accounts in various external domains, lying in the structural and functional characterizability of the phenomena involved. With phenomena such as learning, life, and the weather, the relevant explananda can all be characterized in structural and functional terms; given the causal closure of the physical, one should expect a physical account of this structure and function. As we have seen, consciousness cannot be characterized in these terms, so there is little reason to expect an explanation to be similar in kind.

Indeed, we saw in Chapter 2 that given the nature of our epistemic access to external phenomena, we should *expect* a materialist account to succeed. Our knowledge of these phenomena is physically mediated, by light, sound, and other perceptual media. Given the causal closure of the physical, we should expect phenomena that we observe by these means to be logically supervenient on the physical—otherwise we would never know about them. But our epistemic access to conscious experience is of an entirely different kind. Consciousness is at the very center of our epistemic universe, and our access to it is not perceptually mediated. The reasons for expecting a materialist account of external phenomena therefore break down in the case of consciousness, and any induction from those phenomena will be shaky at best.

Third, many have preferred materialism in order to *take science seriously*. It has been thought that a dualist view would challenge science on its own grounds. According to Churchland (1988), "dualism is inconsistent with evolutionary biology and modern physics and chemistry".[17] But this is quite false. Nothing about the dualist view I advocate requires us to take the physical sciences at anything other than their word. The causal closure of the physical is preserved; physics, chemistry, neuroscience and cognitive science can proceed as usual. In their own domains, the physical sciences are entirely successful. They explain physical phenomena admirably; they simply fail to explain conscious experience.

A fourth motivation to avoid dualism, for many, has stemmed from various spiritualistic, religious, supernatural, and other anti-scientific overtones. But those are quite inessential to the view. On the view I advocate, consciousness is governed by natural law, and there may eventually be a reasonable scientific theory of it. There is no *a priori* principle that says that all natural laws will be physical laws. To deny physicalism is certainly not to deny naturalism. Dualism expands our view of the world, but it need not invoke the forces of darkness.

In a related concern, many have thought that to accept dualism would be to give up on explanation. In the words of Dennett (1991, p. 37): "given the way that dualism wallows in mystery, accepting dualism is giving up". This may have been a feature of some dualist views that have been advocated, but it is inessential. I hope that the rest of this work will demonstrate that significant progress toward an explanation of consciousness is possible within a dualist framework.

In short, very few of the usual reasons for rejecting dualism have any force against the view I am advocating. The main residual reason to reject dualism may simply lie in all the bad connotations of the term, and the fact that it goes against what many of us have been brought up to expect. But once we see past these associations, we see that there is no reason why dualism cannot be a reasonable and palatable view. Indeed, I think that the position I have outlined is one that those who think of themselves as materialists, but who want to take conscious experience seriously, can learn to live with and even to appreciate.

Indeed, mine is a view that many who think of themselves as "materialists" already

[17]Churchland suggests a number of other reasons to reject dualism: the systematic dependence of mental phenomena on neurobiological phenomena, modern computational results that suggests that complex results can be achieved without a non-physical homunculus, and a lack of evidence, explanation, or methodology for dualism. The first two reasons do not count against my view; arguments for dualism have already been presented; and dualist explanation and methodology will be illustrated in the remainder of this work.

implicitly share—those who hold that consciousness "arises" from the physical, for instance. All I have done is draw out the ontological implications of this, so that we know what we are committed to. For reasons like this, many long-time dualists may find me a dubious ally, seeing my view as all too materialistic for their tastes. If so, then so be it. Ideally, it is a view that takes the best of both worlds and the worst of neither.

This dualism, then, requires us to give up little that is *important* about our current scientific world-view. It merely requires us to give up a dogma. Otherwise, the view is merely a *supplement* to our scientific-world view: a necessary broadening in order to bring consciousness within its scope. Credo: if this is dualism, then we should learn to love dualism.

4.7 Toward a nonreductive theory of consciousness

As we saw a moment ago, it is sometimes supposed that non-physicalist views must give up on the possibility of further explanation of consciousness. Such theories are variously described as "spiritualistic", "non-naturalistic", "unscientific", or "requiring miracles".[18] These charges are unfounded. On the view I have put forward, consciousness is certainly part of the natural order—it is just that our view of the natural order has to be expanded. The facts about nature do not include just the facts about physics and their consequences; and the laws of nature do not include just the laws of physics and their consequences. There are also the facts about consciousness, and there are laws of nature linking consciousness with the physical.[19]

Certainly these laws—call them 'supervenience laws', or 'psychophysical laws'—need not be miraculous, any more than the laws of physics are miraculous. There will be something brute about them, it is true. At some level they will have to be taken as true and not further explained. But the same thing holds of the laws of physics. The ultimate laws of nature will always seem arbitrary at some level. It is precisely this that makes them laws of nature rather than laws of logic.

The mere fact that the connection between consciousness and the physical will at some level be a brute fact about nature should not prevent us from trying to systematize this connection as far as possible. We need not leave the brute fact in question at the level of 'this physical state produces this conscious state, that brain produces that conscious state, computer X will produce conscious state Y, and so on'. Instead,

[18]Dennett (1991) suggests that dualism would require miracles. Flanagan (1992) uncritically equates naturalism with materialism. The other charges are widespread.

[19]Views like this, on which consciousness and the physical are linked by irreducible laws, are put forward by Honderich 1981 and Grossman 1992.

we can systematize this connection via an underlying explanatory framework, finding the basic underlying laws in virtue of which this connection holds. Physics does not content itself to be a mere mass of observations about the positions, velocities, and charges of various objects at various times; it systematizes these observations, showing how they are consequences of certain underlying laws, and furthermore seeks to make these laws as simple, as unarbitrary, as possible (witness the quest for a unified field theory). Precisely the same should hold for a theory of consciousness. We should seek to systematically explain the supervenience of consciousness on the physical in terms of the simplest possible set of underlying laws. At some level these laws will be brute, but we can seek to reduce their brutality as far as possible.

It will no doubt be objected that this theory cannot explain *what consciousness is*, or *why consciousness exists*, and in a sense that is true. While such a theory might in a sense explain the existence of conscious experience by giving an account of the laws by which consciousness arises from the physical, it cannot explain why those laws exist, and even more fundamentally it cannot explain what it *is* that those laws are about: conscious experience. Such a theory has to take some aspect of consciousness as primitive, a given of the theory.

While it would be nice to have a theory that explained consciousness fully in terms of something else, this is just what we are not going to get, if the arguments earlier in this chapter are correct. There is an obvious analogy with physics. At the bottom level, physics does not tell us what the physical *is*, or why matter *exists*. It has to take something as primitive, and work with it. A theory of consciousness will have to do the same thing. We cannot expect a theory of consciousness to explain what consciousness *is*, any more than we can expect physics to explain what matter *is*. We just have to take it as given, and work on systematizing how it relates to everything else.

Nevertheless, it may be possible that our ultimate theory will not take 'consciousness' as a primitive. It might be the case that consciousness can be explained in terms of something more basic, as long as this more basic kind is not wholly physical. The same problem will arise, of course, for the more basic notion, and something will eventually have to be taken as primitive, so we are not getting something for nothing. But *a priori*, there are many degrees of freedom in the way the explanation might go.

Two problems are sometimes raised for dualist views of the kind I am suggesting. The first: how do consciousness and the physical interact? The answer to this is straightforward: they interact by virtue of psychophysical laws of the kind I have described. There is a system of laws that requires that a given physical configuration will be accompanied by a given conscious experience, just as there is a law that dictates that a given physical object will gravitationally affect other physical objects in a certain way.

Somebody might object that this does not tell us what the *connection* is by which the law "works". But the search for such a connection is misguided. Even with physical laws, such as the law of gravitation, we cannot find any "connection" that does the work. Things simply happen in accordance with the law. Hume showed us that the quest for such ultimate connections is fruitless. If there are indeed such connections, they are entirely mysterious, both in the case of physical laws and psychophysical laws. The laws connecting consciousness to the physical pose no *special* problems here. The search for a nexus is as futile in either case.

A second objection is that dualist views leave it mysterious how consciousness could ever have arisen in evolution. Somewhere along the line, some entirely new things just popped into existence, as if by magic. But on the view I am suggesting this is no problem. Like the fundamental laws of physics, psychophysical laws are eternal, having existed since the beginning of time. It may be that in the early stages of the universe there was nothing that satisfied the physical antecedents of the laws, so there was no consciousness (although this depends on the precise nature of the laws); but in any case as the physical universe developed, particularly through evolution, it came about that certain physical systems evolved that satisfied these antecedents. When these systems evolved, conscious experience came into being by virtue of the psychophysical laws in question. Given that psychophysical laws exist, as any naturalistic dualist view must surely require, and that these are timeless, the evolution of consciousness poses no particular problem.

There is one enormous problem for a theory of consciousness that does not arise for a theory of physics, and that is the paucity of data. If consciousness does not logically supervene on the physical, then we have direct access to data about consciousness only in our own case. We cannot simply run experiments, setting up various physical configurations and seeing what conscious states arise, as we have access only to the physical properties of the experimental configuration—unless we run the experiments on ourselves! Worse, it might be argued, even in our own case we only have direct access to our conscious states *right now*. Access to other states is possible only through memory, which might be mistaken.

It therefore seems that theories of consciousness will be radically underconstrained. Almost any theory would seem to be compatible with the data, from solipsism (the theory that only I am conscious) to panpsychism (the theory that everything is conscious), via such intermediate theories such as biochemicalism (consciousness arises only from biochemical organisms) and computationalism (consciousness arises from anything with the right computational organization), including along the way such bizarre theories as the theory that people are only conscious in odd-numbered years (right now, it is 1993). We cannot just poke inside a fly's head to see if it is conscious, so will this not remain an eternal mystery?

This is certainly a problem—it is what makes studying consciousness more difficult than studying physics—but I do not think that it is an insurmountable one. All that the above considerations show is that all kinds of theories are *logically compatible* with the data—it does not make them *plausible*. For instance, solipsism seems to be an entirely implausible theory, even though one cannot prove that it is wrong. All kinds of plausibility constraints play a role in shaping our theories, over and above the empirical evidence, in all kinds of domains. Consider, for example, our acceptance of the theory of evolution, as opposed to the theory whereby the world was created 50 years ago with memories and fossil record intact. Or consider the acceptance of nondeterministic theories of quantum mechanics over empirically equivalent but highly jury-rigged hidden-variables theories. Empirical evidence is not all we have to go on in theory-formation; there are also principles of plausibility, simplicity, and aesthetics, among other considerations.

When it comes to consciousness, empirical evidence is hard to find, but a few principles of plausibility go a long way. Consider for example the following exceedingly plausible constraints that a theory of consciousness should satisfy.

(1) Fundamental laws are homogeneous in space-time.

(2) Conscious experience depends only on the internal physical state of an organism.

(3) People's reports of their conscious experiences, by and large, accurately reflect the contents of those experiences. (Note that "reflect" need not mean "are caused by".)

(4) People's memories of conscious experiences are not radically mistaken, by and large.

One cannot prove or empirically test these constraints, but they are nevertheless very likely to be true. If they were not true, the world would be an unreasonable place. In developing scientific theories, one has to assume that the world is a reasonable place, where planets do not suddenly pop into existence in 1942, and where hidden-variable laws are not jury-rigged to reproduce the predictions of a simpler quantum theory; otherwise, anything goes. And these principles already provide strong constraints on a theory of consciousness. Solipsism and the odd-numbered-year theory are immediately ruled out, for instance. As a bonus, these principles provide immediate purchase for experimental research on consciousness—to determine whether a person is having a particular kind of conscious experience, just ask them! Of course this method cannot be overextended, and is not so useful for finding out about the conscious experiences of dogs, say, but it is a start.

Furthermore, once one accepts a certain reliability of memory, one is able to make use of systematic correlations between physical/functional properties and conscious experience in one's own case (without the memory assumption, it might have been argued that these correlations were illusory). Given these regularities, one can search for principles that subsume them as simply as possible, by a kind of inference to the best explanation, and one can thus infer a coherent system of psychophysical laws. One might think that this system would still be underdetermined, but principles of simplicity, coherence, and universality go a long way toward constraining it. There is also a significant role that can be played by thought-experiments, as we will see in Chapter 7.

This is just a glimpse of what plausibility arguments can provide us in developing theories of consciousness. In later chapters, I will develop this sort of argument much further, moving toward a far more specific theory of consciousness. Of course, one will never be able to conclusively *test* one's theories, as one can with more common varieties of scientific theories, but that is the way things are. We take the materials we have, and we work with them.

The mode of this work will now shift back from ontology to explanation. I will take it as given that consciousness supervenes on the physical, and will not worry as much about whether the mode of that supervenience is nomic, metaphysical, or conceptual. Instead I will investigate what the relevant principles of supervenience *are*. Consciousness certainly arises from the physical, one way or the other. Most of the remainder of this work will be devoted to the question of *how* it so arises, and in particular to the question: in virtue of what physical properties does consciousness arise? Chapters 6 and 7 will be devoted to this question in different ways, as will further work summarized in an appendix. I will not offer a fully-developed theory of consciousness, but I hope to put some strong constraints on what such a theory should look like.

4.8 Appendix: Some other views

There are various other views on consciousness to be found in the literature that do not explicitly fall into the framework I outlined in 4.6. I will briefly discuss some of these views, noting how they stand with respect to the framework, and why they are problematic.

1. *Biological materialism.* A common view is that only beings of a certain biological type have conscious experience. Such a view is advocated by Hill (1991) and Searle (1992), among others. Those who hold this view generally hold that there could be

a system with the same functional organization as me but which is not conscious, as it is made of the wrong material. This view is often put forward as a version of materialism (e.g. by Hill 1991 and Searle 1992), but it can be straightforwardly seen to be incompatible with materialism.

To see this, note that once we have admitted the (empirical or logical) possibility of an unconscious silicon functional isomorph of me, we must surely admit the logical possibility of an unconscious *biological* functional isomorph of me, and in the limit, an unconscious physical replica of me: that is, a zombie. There is no more of a *conceptual* link from neurophysiology to conscious than from silicon to consciousness.[20] If a silicon zombie is conceivable, a biological zombie is equally conceivable. Once this is accepted, then it becomes clear that consciousness is a further fact over and above the physical, and that materialism is false.

It seems to me that those who put forward this view should describe themselves as dualists, if they are to be consistent. An alternative is that this position might be combined with alternative (vi) above, so that there is a brute "metaphysically necessary" but not a logically necessary connection from biochemistry to consciousness, but then the view inherits all the problems of that position.

2. *Psychofunctionalism.* On this view, mental properties are identified with functional properties not *a priori* but *a posteriori*, on the basis of their roles in a mature empirical psychology (see Block 1980). If this view applied to phenomenal properties, phenomenal notions would have the same posterior intensions as functional notions, despite a difference in prior intension. The view is therefore vulnerable to the arguments against the first "metaphysical necessity" objection above—we simply focus on the property introduced by the prior intension—and so cannot do any work against materialism.[21]

Advocates of this view have often ignored the role of concepts in fixing reference via prior intensions. Even when we have some complex scientific theory with 'belief' as a theoretical term, there will still be a story to tell about why *that* sort of state qualifies as a *belief*, rather than as a desire or something else entirely. That story will be entirely conceptual. Presumably the reference-fixing intension for 'belief' will itself be functional—something like, 'the state that plays the most belief-like role within the theory', where 'belief-like' is cashed out according to our prior concept.

[20] Searle admits the logical possibility of zombies, and in fact holds that there is merely a causal connection between the microphysical and conscious experience, but holds that this is compatible with materialism, as we saw earlier. Hill tries to avoid the possibility by an appeal to rigid designators, but we have seen that this strategy does not help.

[21] Some have supposed that psychofunctionalism only requires that functional and phenomenal properties coincide with *nomic* necessity (e.g. Hill 1991); but we have seen that this is too weak to qualify as materialism.

The prior intension for phenomenal properties, however, is surely not functional. If the theory claims that it is, then it falls into all the problems with misanalysis of a concept that we attributed to analytic functionalism, problems that this view hoped to avoid. In any case, whatever the prior intension, then all the problems for materialism will arise for it, as we have seen. To concentrate on posterior intensions is simply to sweep the problems under the rug.[22]

3. *Physicalist-functionalism.* On this popular view, advocated by Shoemaker (1982) among others, the property of having a conscious experience is functional, but the property of having some *specific* conscious experience—a certain kind of red sensation, say—is neurophysiological. This view is generally put forward as a consequence of the empirical possibility that two functionally isomorphic systems could have spectra inverted with respect to each other. Where one has red sensations, the other has blue sensations, and so on.

Once we have accepted that it is empirically or logically possible that my functional isomorph could have inverted experiences, however, it is equally clear that it is logically possible that my *physical* isomorph could have the same. As before, there is no more of a *conceptual* connection from neurophysiology to a particular experience than there is from silicon. It follows that the physical facts do not logically determine all the facts, and that physicalism is false. (Alternatively, this view might be combined with the brute metaphysically necessary connection of alternative (vi), but this would inherit all the problems of that view.)

This view is often put forward as an *a posteriori* identification of phenomenal properties with neurophysiological properties. As such, it is vulnerable to all the usual problems with such *a posteriori* identification, as well as to the argument above. Those who advocate this view should stick to an across-the-board functionalism. The latter may not be satisfactory, but at least it is consistent. See White (1986) for an in-depth critique of physicalist-functionalism along these lines.

4. *Anomalous monism.* On this view, each mental state is token-identical to a physical state, but there are no strict psychophysical laws. Anomalous monism was put forward by Davidson (1970) as an account of intentional states such as beliefs and desires rather than as an account of phenomenal states, but it might still be thought relevant for two reasons: first, it offers an *a priori* argument for physicalism

[22]Another problem with this view is that it implies a kind of chauvinism, by giving an extra weight to *human* psychology in deciding what counts as a belief, say. See Shoemaker 1981 for an excellent critique, although see Clark 1986 for a response. It seems more plausible that even for a functional notion like belief, the prior and posterior intensions will coincide. Otherwise, we get into situations where we and our Twin Earth counterparts mean different things by 'belief', despite our prior concepts being identical.

based simply on the causal interaction (even a one-way interaction) between physical and mental states, and second, it denies the psychophysical laws that my view requires. The arguments about intentional states are adaptable to arguments about phenomenal states without too much difficulty.

To see that nomic supervenience is not threatened by Davidson's arguments, note that nothing in his arguments counts against the existence of *pointwise* laws of the form 'if a system is in maximally specific physical state P, it is in (maximally specific) mental state M'. Indeed, Davidson endorses the supervenience of the mental on the physical, which seems to have the existence of such laws as a consequence, upon a natural interpretation (see Kim 1985 for discussion). I think the most charitable interpretation of Davidson reads him not as denying such pointwise laws, but merely as denying more interesting *typewise* laws connecting mental states to physical states under broad types such as those of folk psychology (see Childs 1993 for a discussion along these lines). Certainly none of his arguments from the holism of the mental will count against pointwise laws, while they may count against a sort of typewise law.

If so, then nomic supervenience is not threatened. It also follows that the argument for token identity cannot go through. This argument relied on there being *no* strict laws connecting the physical and the mental to support a causal connection between these, so that an identity would be required instead. As it is, even a strict pointwise law is sufficient to underwrite the kind of connection that we need, from physical states to phenomenal states. So dualism is not threatened either.

5. *Emergent causation.* Many have been concerned to reject a reductive account of consciousness while nevertheless ensuring it a more direct causal role than the view I have presented can provide. A popular way to do this has been to argue for *emergent causation*—the existence of new sorts of causation in physical systems when matter is organized in certain complex ways. Sperry (1969; 1992) has advocated such a view, arguing that consciousness is an emergent property of complex systems that itself plays a causal role. Sellars (1978; see also Meehl and Sellars 1958) was also sympathetic with such a view, advocating its possibility under the name of "physicalism$_1$". This is the view that new laws of physical causation might come into play in certain sorts of systems, such as those made of protoplasm or those supporting sentient beings, as opposed to "physicalism$_2$", the view that the basic physical principles found in inorganic matter apply across the board.

It is important not to confuse this with the "innocent" view of emergent causation (found in complex systems theory and artificial life, for example), on which low-level laws yield qualitatively novel behavior via complex interaction effects. On the more radical view, there are new fundamental principles at play that are not consequences of low-level laws. A view like this was also held by the British emergentists such as Alexander (1920); see McLaughlin 1992 for a careful explication of their views and a critique.

There are two problems with the view. The first is that there is no evidence that such emergent principles of causation ever come into play. As far as we can tell, all causation is a consequence of low-level physical causation. This view would require a kind of downward causation that would lead to breaches in the principles of low-level causation that apply elsewhere, and there is no evidence for such breaches.

Still, proponents might argue that there *must* be this sort of emergent causation if an irreducible consciousness is to play a causal role; certainly, this has been the most common motivation for such a view. This leads us to the more important second problem, which is that upon close examination, such a view leaves consciousness as epiphenomenal as ever. To see this, note that nothing in the story about emergent causation requires us to invoke *phenomenal* properties anywhere. We can tell the whole story in terms of physical properties, noting that certain complex configurations of physical properties yield certain physical consequences, and so on. There will still be a possible world that is physically identical but that lacks consciousness entirely. Perhaps phenomenal properties *correlate* with causally efficacious configurations on this view, but this no more allows them an independent causal role than my view does; conversely, insofar as this view allows phenomenal properties a causal role, so does mine. In fact, this view is best seen as a version of my view, with consciousness nomically supervenient on the physical. It is modified by the addition of new laws of emergent causation, but these simply complicate matters somewhat, rather than changing anything fundamental.

6. *Mysterianism.* Those unsympathetic to reductive accounts of consciousness often hold that consciousness may remain an eternal mystery. Nagel (1974) has suggested such a view, and it has been developed by McGinn (1991). Jackson (1982) also suggests that conscious experience may be beyond our understanding, just as sea slugs cannot understand most of the universe.

Such a view can be tempting in view of all the problems that consciousness poses, but it is premature. To say that there is no *reductive* explanation of consciousness is not to say that there is no explanation at all. In particular, an explanation of the principles in virtue of which consciousness nomically supervenes on the physical might provide an enlightening theory of consciousness even on a nonreductive view.

McGinn (1991) argues that (a) there is a necessary connection between brain states and conscious states (otherwise the emergence of consciousness would be a miracle) and that (b) we can never know what this connection is. He does not explicitly specify the grade of necessity in question. The unqualified usage suggests that he has a logically or metaphysically necessary connection in mind; but his *argument* for a necessary connection only establishes a nomic connection. Certainly a contingent nomic connection between consciousness and the physical is no more miraculous than any contingent law. This is far more satisfactory than a view on which there is a logically or metaphysically necessary connection of a kind that we cannot understand; such a position would seem to collapse into the "brute metaphysical necessity" view canvassed earlier.

McGinn argues that we could not know the physical properties in virtue of which consciousness arises, but the argument seems weak to me. Rather than rebut it directly, I will let the following chapters speak for themselves. In these chapters, I move some distance toward isolating the relevant class of physical properties in virtue of which consciousness arises. In this way, we can see that non-reductivism about consciousness need not lead to pessimism.

Chapter 5

The Paradox of Phenomenal Judgment

5.1 Consciousness and cognition

In previous chapters, the distinctions and divisions between consciousness and cognition have been stressed above all else. Consciousness is mysterious; cognition is not. Consciousness is ontologically novel; cognition is an ontological free lunch. Cognition can be explained functionally; consciousness resists such explanation. Cognition is governed entirely by the laws of physics; consciousness is governed in part by independent psychophysical laws.

While the focus on these distinctions has been necessary in order to come to grips with the many subtle metaphysical and explanatory issues surrounding conscious experience, it may encourage a misleading picture of the mind. On this picture, consciousness and cognition are utterly alienated from each other, living independent lives. One might get the impression that a theory of consciousness and a theory of cognition will have little to do with one another.

This picture is radically misleading. Our mental life is not alienated from itself in the way that this picture suggests. There are deep and fundamental ties between consciousness and cognition. On one side, the contents of our conscious experience are closely related to the contents of our cognitive states. Whenever one has a green sensation, individuated phenomenally, one has a corresponding green *perception*, individuated psychologically. On the other side, much cognitive activity can be centered on conscious experience. We know about our experiences, and make judgments about them; as I write this, a great deal of my thought is being devoted to consciousness. These relations between consciousness and cognition are not arbitrary and capricious, but systematic.

It turns out that insights about the nature of cognition can provide much of the basic material for a theory of consciousness. Of course a theory of cognition cannot do all the work that needs to be done, as we have seen, but this does not mean that it cannot play a major role. After all, it is through cognition that we get a handle on consciousness in the first place. A thorough investigation of the links between consciousness and cognition can provide the purchase we need to constrain a theory of consciousness in a significant way, and perhaps to ultimately produce an account of consciousness that neither mystifies nor trivializes the phenomenon.

Previous chapters have sundered the phenomenal aspects of mind from the psychological. In this chapter, I begin the task of drawing them together into a unified picture of mind, a task that will be continued in the remainder of this work.

5.2 Phenomenal judgments

The primary nexus of the relationship between consciousness and cognition lies in our *judgments about consciousness.* Our conscious experience does not reside in an isolated phenomenal void. We are aware of our experience, we form judgments about it, and we are led to make claims about it. When I have a red sensation, I often form a belief that I am having a red sensation, which can issue in a verbal report. At a more abstract level, when one stops to reflect on the mysteries that consciousness poses—as I have been doing throughout this work—one is making judgments about consciousness. These judgments I will call *phenomenal judgments*, not because they are phenomenal states themselves, but because they are about phenomenology. (Compare: moral judgments, economic judgments, and so on.)

Phenomenal judgments are often reflected in *claims* about consciousness: verbal reports of the contents of those judgments. At various times, people make claims about consciousness ranging from "I have a throbbing pain now" through "LSD gives me bizarre color sensations" to "The problem of consciousness is utterly baffling". These claims and judgments are intimately related to our phenomenology, but they are ultimately part of our psychology. Verbal reports are behavioral acts, and are therefore susceptible to functional explanation. In a similar way phenomenal judgments are themselves cognitive acts, and fall within the domain of psychology.

I have argued in earlier chapters that beliefs should be understood as functional states, characterized by their causal ties to behavior and the environment. We saw then that this view is not universally accepted, and that some hold that phenomenal qualities can be partly constitutive of belief, or of belief contents. For beliefs about consciousness, the functional view is likely to be particularly controversial: if any beliefs are dependent on conscious experience, beliefs about consciousness are the most likely candidates. I will therefore adopt the less loaded label 'judgment' for the functional states in question, and will leave open the question of whether a judgment about consciousness is all there is to a belief about consciousness. We can think of a judgment as what is left of a belief after any associated phenomenal quality is subtracted.

That there are purely psychological states that qualify as these judgments should not be a controversial matter. For a start, the disposition to make certain verbal reports is a psychological state; at the very least, we can use the label 'judgment' for this disposition. Moreover, whenever I form a belief about my conscious experience, there are all sorts of accompanying functional processes, just as there are with any belief. These processes underlie the disposition to make verbal reports, and all sorts of other dispositions. If one believes that LSD produces bizarre color sensations, the accompanying processes may underlie a tendency to indulge in or to avoid LSD in future, and so on. We can use the term 'judgment' as a coverall for the states or processes that play the causal role in question. At a first approximation, a system judges that P if it tends to respond affirmatively when queried about P, to behave in a manner appropriate for P given its other beliefs and desires, and so on.

The judgments in question can perhaps be understood as what I and my zombie twin have in common. My zombie twin does not have any conscious experience, but he certainly *claims* that he does, and he makes the same detailed verbal reports as I do. I think it is natural to say that my zombie twin *believes* he has conscious experience, and that he has the same detailed beliefs that I do (modulo the fact that his beliefs are about *him*, whereas mine are about *me*). Rather than simply assuming this substantive thesis about beliefs, however, we can simply note that in the matter of belief, there is much that my zombie twin and I have in common. There is *some* not unreasonable sense of 'belief' in which my zombie twin believes what I believe. This shared aspect we will call 'judgment'. My zombie twin and I make the same judgments; it is just that many of his judgments—in particular those concerning conscious experience—are false.

Judgments related to conscious experience fall into at least three groups. There are what I will call *first-order*, *second-order*, and *third-order* phenomenal judgments. (I will usually drop the qualifier and speak of 'first-order judgments' and so on. It should be understood that the judgments in question are always phenomenal judgments.)

First-order judgments are the judgments that go along with conscious experiences, concerning not the experience itself but the *object* of the experience. When I have a red sensation—upon looking at a red book, for instance—there is generally an explicit or implicit judgment "there is something red". When I have the experience of hearing a musical note, there is an accompanying psychological state concerning that musical note. It seems fair to say that any object that is *consciously experienced* is also *cognitively represented*, although there is more to say about this (I will discuss this further in the next chapter). Alongside every conscious experience there is a content-bearing cognitive state. This cognitive state is a first-order judgment.

First-order judgments are not strictly *about* consciousness; rather they are *parallel* to consciousness. The contents of these judgments do not generally concern conscious experiences, but objects and properties in the environment (or even in the head). Still, these judgments are intimately bound up with what is consciously experienced, and the link between these judgments and conscious experience itself is an important part of the story that needs to be told about the relation between consciousness and cognition.

First-order judgments need not be explicit or occurrent judgments, or judgments that we endorse upon reflection. For instance, my cognitive representation of a red object may not be in the foreground of my thought processes, if my attention is directed elsewhere. Still, as long as it is in my perceptual field, it is represented, and I have some psychological state bearing the content that the object is red. This state is a first-order judgment. In a similar way, when we see an optical illusion we may form a first-order perceptual judgment that one object is larger than another, even if on reflection we judge that they are the same size. The first-order judgments with which we will be most concerned are perhaps best regarded as the immediate products of various perceptual and introspective processes, before these are rationally integrated into a coherent whole; they might also best be regarded as implicit rather than explicit judgments. It is these cognitive states that most directly parallel the contents of consciousness.

Second-order judgments are more straightforwardly judgments about conscious experiences. When I have a red sensation, I sometimes notice that I am having a red sensation. I judge that I have a pain, that I experience certain emotional qualities, and so on. In general, it seems that for any conscious experience, one at least has the capacity to judge that one is having that experience.

One can also make more detailed judgments about conscious experiences. One can note that one is experiencing a particularly vivid shade of purple, or that some pain has an all-consuming quality, or even that some green after-image is the third such after-image one has had today. Apart from judgments about specific conscious experiences, second-order judgments also include judgments about particular *kinds* of conscious experiences, as when one notes that some drug produces particularly intense sensations, or that the tingle one gets before a sneeze is particularly pleasurable.

What I will call *third-order* judgments are judgments about conscious experience as a type. These go beyond judgments about particular experiences. We make third-order judgments when we reflect on the fact that we have conscious experiences in the first place, and when we reflect on their nature. I have been making third-order judgments throughout this work. A typical third-order judgment might be "Consciousness is baffling; I don't see how it could be reductively explained". Others include "Conscious experience is ineffable", and even "Conscious experience does not exist".

Third-order judgments are particularly common among philosophers, obviously, and among those with the tendency to speculate on the mysteries of existence. It is possible that many people go through life without making any third-order judgments. Still, such judgments occur in a significant class of people. The very fact that people make such judgments is something that needs explanation.

To help keep the distinctions in mind, the various kinds of judgments related to consciousness can be represented by the following:

First-order judgment: *There is a red object.*
Second-order judgment: *I'm having a red sensation now.*
Third-order judgment: *Aren't sensations mysterious?*

5.3 The paradox of phenomenal judgment

The existence of phenomenal judgments poses an obvious tension for a non-reductive theory of consciousness. The problem is this: We have seen that consciousness itself cannot be reductively explained. But phenomenal judgments lie in the domain of psychology, and in principle should be reductively explainable by the usual methods of cognitive science. There should be a physical or functional explanation of why we are disposed to make the *claims* about consciousness that we do, for instance, and even of why we *think* the things we do about conscious experience, if thinking is understood appropriately. It then follows that our claims and judgments about consciousness can be explained in terms quite independent of consciousness. More strongly, it seems that consciousness is *explanatorily irrelevant* to our claims and judgments about consciousness.[1] This result I will call the *paradox of phenomenal judgment*.

I wish to be circumspect about claiming that consciousness is *causally* irrelevant. That is a subtle question that I think is still open; it may depend on the nature of causation itself. The *explanatory* irrelevance of consciousness is clearer. It seems relatively straightforward that a physical explanation of behavior can be given, neither appealing to nor implying the existence of consciousness.

When I say, in conversation "Consciousness is the most mysterious thing there is", that is a behavioral act. When I wrote in an earlier chapter "Consciousness cannot be reductively explained", that was a behavioral act. When I comment on some particularly intense purple qualia that I am experiencing, that is a behavioral act. Like all behavioral acts, these are in principle explainable in terms of the internal causal organization of my cognitive system. There is some story about firing patterns in neurons that will explain why these acts occurred; at a higher level, there is probably a story about cognitive representations and their high-level relations that will do the relevant explanatory work. We certainly do not know the details of the explanation now, but if the physical domain is causally closed, then there will be some reductive explanation in physical or functional terms.

In giving this explanation of my claims in physical or functional terms, we will never have to invoke the existence of conscious experience itself. The physical or functional explanation will be given independently, applying equally well to a zombie as to an honest-to-goodness conscious experiencer. It therefore seems that conscious experience is irrelevant to the explanations of phenomenal claims and irrelevant in a similar way to the explanation of phenomenal judgments, even though these claims and judgments are centrally concerned with conscious experience!

It might be argued that for *any* high-level property that might be thought relevant in explanation, there will be a low-level explanation that does not invoke the existence of that property. One could argue that beliefs are explanatorily irrelevant, as one can give neurophysiological explanations of actions that never once mention a belief; one could even argue that temperature is explanatorily irrelevant in physics, as explanatory appeals to temperature can in principle be replaced by a molecular account. (Kim 1989 calls this the problem of *explanatory exclusion.*) This might suggest that consciousness is no worse off than any other high-level property when it comes to explanatory irrelevance. If consciousness is on a par with temperature, belief, and species-membership, this is not bad company to be in.

We have seen, however, that high-level properties such as belief and temperature, and fitness are all *logically supervenient* on the physical. It follows that when one gives an explanation of some action in neurophysiological terms, this does not make belief explanatorily irrelevant. Belief can *inherit* explanatory relevance by virtue of its logically supervenient status. When we explain a man's desire for female companionship in terms of the fact that he is male and unmarried, this does not make the fact that he is a bachelor explanatorily irrelevant! The general principle here is that when two sets of properties are *conceptually* related, the existence of an explanation in terms of one set does not render the other set explanatorily irrelevant. In a sense, one of the explanations can be a retelling of the other, due to the conceptual relation between the terms involved.

When we tell a story about the interaction of beliefs, there is a sense in which we are retelling the physical story at a higher level of abstraction. This higher level will omit many details from the physical story, and will therefore often make for a much more satisfying explanations (all those details may have been irrelevant clutter), but it is nevertheless logically related to the lower-level story. The same goes for temperature and species-membership. These high-level properties are no more rendered explanatorily irrelevant by the existence of a low-level explanation than the velocity of a billiard ball is rendered explanatorily irrelevant by the existence of molecular processes within the ball. In general, the high-level properties in question will constitute a more parsimonious *redescription* of what a low-level explanation describes. One might say that even a low-level description will often *implicitly* involve high-level properties, by virtue of their logically supervenient status, even if it does not invoke them explicitly. Where there is logical supervenience, there is no problem of explanatory irrelevance.

The problems with consciousness are much more serious. Consciousness is not logically supervenient on the physical, so we cannot claim that a physical or functional explanation implicitly involves consciousness, or that consciousness inherits explanatory relevance by logically supervening on the properties involved in such an explanation.[2] A physical or functional explanation of behavior is independent of consciousness in a much stronger sense. It can be given in terms that do not even *imply* the existence of conscious experience. Consciousness seems to be quite independent of what goes into the explanation of our claims and judgments about consciousness.[3]

To see the problem in a particularly vivid way, think of my zombie twin in the universe next door. He talks about conscious experience all the time—in fact, he seems obsessed by it. He spends ridiculous amounts of time hunched over a computer, writing chapter after chapter on the mysteries of consciousness. He often comments on the pleasure he gets from certain sensory qualia, professing a particular love for deep greens and purples. He frequently gets into arguments with materialists, arguing that their position cannot do justice to the realities of conscious experience.

And yet he has no conscious experience at all! In his universe, the materialists are right and he is wrong. Most of his claims about conscious experience are utterly false. But there is certainly a physical or functional explanation of why he makes the claims he makes. After all, his universe is fully law-governed, and no events therein are miraculous, so there must be *some* explanation of his claims. But such an explanation must ultimately be in terms of physical processes and laws, for these are the *only* processes and laws in his universe.

Now my zombie twin is only a logical possibility, not an empirical one, and we should not get *too* worried about odd things that happen in logically possible worlds. Still, there is room to be perturbed by what is going on. After all, any explanation of my twin's behavior will equally count as an explanation of *my* behavior, as anything that is going on within him is also going on within me. The explanation of *his* claims obviously does not depend on the existence of consciousness, as there is no consciousness in his world. It follows that the explanation of my claims is also independent of

[2]Note that the position whereby consciousness supervenes on the physical by a brute, inexplicable necessary connection will be no better off here. On such an account, consciousness will certainly by *explanatorily* irrelevant to behavior, although there may be some sense in which it is causally relevant. It is conceptual connections that are give explanatory relevance.

[3]This is not to say that one can *never* appeal to conscious experience in the explanation of behavior. It is perfectly reasonably to explain the fact someone's withdrawal from the pain by the fact that they experienced pain. After all, even on the non-reductionist view there are still lawful regularities between experience and subsequent behavior. Such regularities will ultimately be dependent on regularities at the physical level, however. For any explanation of behavior that appeals to a pain sensation, there will be a corresponding explanation in purely physical/functional terms—say, in terms of psychological pain, or pain perception—that does not invoke the experience. In a sense, the validity of the experience-involving explanation will be implicitly dependent on the validity of this explanation, together with the nomic link between physical/functional properties and properties of experience.

the existence of consciousness.

To strengthen the sense of paradox, note that my zombie twin is himself engaging in reasoning just like this. He has been known to lament the fate of *his* zombie twin, who spends all his time worrying about consciousness despite the fact that he has none. He worries about what that must say about the explanatory irrelevance of consciousness in his own universe. Still, he remains utterly confident that consciousness exists. But all this, for him, is a monumental waste of time. There *is* no consciousness in his universe—in his world, the eliminativists are right, and he is wrong. Despite the fact that his cognitive mechanisms function in the same way as mine, *his* judgments about consciousness are quite deluded (although in a sense there is nobody "in there" to delude).

This paradoxical situation is at once delightful and disturbing. It is not *obviously* fatal to the non-reductionist position, but it is at least something that we need to come to grips with. It is certainly the greatest tension that a non-reductionist theory is faced with, and any such theory that does not at least face up to the problem cannot be fully satisfactory. We have to carefully examine the consequences of the situation, and separate what is merely counterintuitive from what threatens the viability of a non-reductionist view of consciousness.

Nietzsche said "What does not kill us, makes us stronger". It turns out that we can cope with the paradox, and furthermore that facing up to the paradox yields valuable insights about consciousness and its relation to cognition. In this way, a theory of consciousness can be set onto much firmer ground. In particular, dealing with problems raised by phenomenal judgments will lead us to important conclusions about the systematic *coherence* between consciousness and cognition, from which we will be much better placed to build up a detailed theory of consciousness. In different ways, each of the next few chapters will be devoted to investigating this relationship, motivated by considerations about phenomenal judgments. Without taking advantage of the nexus between consciousness and cognition that phenomenal judgments provide, a theory of consciousness would be hard-pressed to get off the ground.

Facing up to the paradox

When it comes to the explanation of *most* of our behavior, the fact that consciousness is explanatorily irrelevant may be counterintuitive, but it is not too paradoxical. To explain my reaching for the book in front of me, we need not invoke my phenomenal *sensation* of the book; it is enough to invoke my *perception* instead. When a concertgoer sighs at a particularly exquisite movement, one might have thought that the experienced quality of auditory sensations might be central to an explanation of this behavior, but it turns out that an explanation can be given entirely in terms of auditory perception, and functional responses to it. Even in explaining why I withdraw my hand from a flame, a functional explanation in terms of the psychological notion of pain (outlined in Chapter 1) will suffice.

In general, it turns out that where one might think that one would need to invoke phenomenal properties in the explanation of behavior, one can usually invoke psychological properties instead. We saw in Chapter 1 that there is a psychological state corresponding to every phenomenal state. Where one might have invoked a sensation, one invokes a perception; where one might have invoked the phenomenal quality of an emotion, one invokes a corresponding functional state; where one might have invoked an occurrent thought, one need only invoke the content of that thought. It is this correspondence between phenomenal and psychological properties that makes the explanatory irrelevance of phenomenal properties not *too* serious a problem in general. It is counterintuitive at first, but it is only counterintuitive. At least for behavior that is not directly concerned with conscious experience, there does not seem to be a pressing need to invoke phenomenal properties in explanation.

It is with our claims and judgments about consciousness that the explanatory irrelevance of conscious experience becomes troubling. True, it may not be especially worrying that consciousness is explanatorily irrelevant to our *first-order* phenomenal judgments—those that are parallel to conscious experience, such as "There is a red object there". It is reasonable that these should be explained purely in terms of perception and other psychological processes; after all, the judgments in question are not directly concerned with conscious experience, but with the state of the world.

For second- and third-order phenomenal judgments, however, explanatory irrelevance seems to raise real problems. It is these judgments that are *about* conscious experience, and that are responsible for our talking about our sensations and for philosophers' worries about the mysteries of consciousness. It is one thing to accept that consciousness is irrelevant to explaining how I walk around the room; it is another to accept that it is irrelevant to explaining why I talk about consciousness! One would surely be inclined to think that the fact that I am conscious will be part of the explanation of why I *say* that I am conscious, or why I *judge* that I am conscious; and yet it seems that this is not so.

After all, part of the explanation of why we claim and judge that there is water will involve the fact that there is indeed water. In a similar way, it seems that the existence of stars and planets is almost certainly explanatorily relevant to our judging that there are stars and planets. As a rule, when we judge truly and reliably that P, the fact that P is true generally plays a central role in the explanation of the judgment. True, for some of our claims and judgments, the objects of those judgments are explanatorily irrelevant to the judgments themselves. Think of religious beliefs, for instance, which might be explained without invoking any gods, or beliefs about UFOs. But these are all probably *false* beliefs, and certainly not instances of *knowledge*. By contrast, we *know* that we are conscious.

Here we are faced with a difficult situation: how can knowledge of consciousness be reconciled with the fact that consciousness is explanatorily irrelevant to phenomenal judgments? If phenomenal judgments arise for reasons independent of consciousness itself, does this not mean that they are unjustified? This, above all, is the central difficulty posed by the paradox of phenomenal judgment, and I will address it at length later in the chapter. If the failure of logical supervenience implies explanatory irrelevance, and if the explanatory irrelevance of P to a judgment that P rules out knowledge of P, then we are faced with not only counterintuitiveness but contradiction. However, we will see later that the consequences need not be so dire.

For now, let us sum up the situation. There are a number of plausible principles, each of which we have good reason to believe in, but which together raise significant tensions.

(1) We know there is conscious experience.
(2) Conscious experience is not logically supervenient on the physical.
(3) The physical domain is causally closed.
(4) Judgments about consciousness are logically supervenient on the physical.

From (3) and (4), it follows that judgments about consciousness can be reductively explained. In combination with (2), this seems to indicate that conscious experience is explanatorily irrelevant to our judgments about consciousness, which leads to an obvious tension with (1).

Some might be tempted to deny (4), but remember that we have *defined* judgments so that they are functional states, logically supervenient on the physical. Now, some might argue that there is no such functional state that remotely resembles what we think of as a judgment; but even so, we can simply retreat to *claims* about consciousness, which are behavioral acts and so more straightforwardly logically supervenient, and which raise the difficulties just as strongly. Even if someone argued that behavioral acts are not purely physical (they might argue that conscious experience is required for something to qualify as a *claim* rather than a noise, or as a claim *about consciousness*), it is still surprising that consciousness is explanatorily irrelevant to the sounds we produce, and to the marks we write, all of which can be systematically interpreted as concerning consciousness. So the situation cannot be escaped this easily.

The status of judgments may nevertheless be a key to the difficulties here, but if so it will not be because of their failure to logically supervene. Rather, it may be that functionally individuated judgments do not capture all that is important about our epistemic situation with respect to consciousness. Perhaps the justification of our belief in consciousness does not lie in the explanatory source of the judgment, but somewhere else: perhaps in a non-causal direct acquaintance with conscious experience. If so, then the explanatory irrelevance of consciousness to our judgments is less in tension with our knowledge of consciousness than it might seem. I will pursue a line like this later.

Other anti-reductionists may be tempted to deny (3). Traditionally, many of those who have repudiated physicalism about consciousness have been drawn to a Cartesian interactionist dualism, thinking that only this can give consciousness the relevance to our action that it surely deserves. Indeed, Elitzur (1989) argues directly from the existence of claims about consciousness to the conclusion that the laws of physics cannot be complete, and that consciousness plays an active role in directing physical processes (he suggests that the second law of thermodynamics might be false). But this is surely premature. It raises all the problems with interactionist dualism that we have seen in earlier chapters, and most importantly it suggests that physicists are wrong in their own domain. While the option should perhaps be left open, interactionist dualism is a last resort.

Others will want to take a reductionist line and deny (1) or (2). I have argued exhaustively for (2) in earlier chapters, so I will not repeat the arguments here. The denial of (1) would lead us back to eliminativism about consciousness, a position that is utterly at odds with the evidence. Still, I will examine a strategy along the lines of denying (1) or (2) in more detail shortly.

It seems to me that the most reasonable attitude to take is to recognize that (1), (2), (3) are all probably true ((4) is true by definition), and to see how they can be reconciled. We *do* know there is conscious experience; the physical domain is almost certainly causally closed; and we have established earlier that consciousness is not logically supervenient on the physical. We have to learn to live with the combination.

5.4 On explaining phenomenal judgments

Given what has gone before, explaining why we say the things we do about consciousness emerges as a reasonable and interesting project for cognitive science. These claims are behavioral acts, and should be as susceptible to explanation as any other behavioral act. Indeed, there should be rich pickings for any cognitive scientist who takes this path. Explaining our claims and judgments about consciousness may be difficult, but it will not be as difficult as explaining consciousness itself. This explanation will not automatically yield an explanation of consciousness, of course, but it may well point us in the right direction. (I will discuss this link between the explanations later in the chapter.)

We can do more than accept the possibility of such an explanation as an intellectual conclusion, derived from the causal closure of physics and the logical supervenience of behavior. There are independent reasons for thinking that phenomenal judgments will be natural concomitants of certain kinds of cognitive processes, and that on reflection one should *expect* such judgments from an intelligent system with a certain design. If so, then the explanation of the claims and judgments may not be as difficult as one might think; they might fall out of some basic principles about cognitive design.

To get some inkling of this, imagine that we have created computational intelligence in the form of an autonomous agent that perceives its environment and has the capacity to reflect rationally on what it perceives. What would such a system be like? Would it have any concept of consciousness and of related notions?

To see that it might, note that on the most natural design such a system would surely have some concept of self—for instance, it would have the ability to distinguish itself from the rest of the world, and from other entities resembling it. It also seems reasonable that such a system would be able to access its own cognitive contents much more directly than it could those of others. If it had the capacity to reflect, it would presumably have a certain direct awareness of its own thought contents, and could reason about that fact. Furthermore, such a system would most naturally have direct access to its perceptual inputs, in the way that we have direct access to ours.

When we asked the system what perception was like, what would it say? Would it say "It's not like anything"? Perhaps it might say "Well, I know there is a red tricycle over there, but I have no idea *how* I know. The information just appeared in my database." Perhaps, but it seems unlikely. A system designed this way would be quite inefficient and unnatural; its access to its own perceptual contents would be curiously indirect. It seems much more likely that it would say "I know there is a red tricycle because I *see* it there." When we asked it how it *knows* that it is seeing the tricycle, the answer would very likely be something along the lines of "I just see it."

It would be an odd system that replied "I know I see it because sensors 78-84 are activated in such-and-such a way." There is no need to give a system such detailed access to its low-level parts. Even Winograd's program SHRDLU (1972) did not have knowledge about the code it was written in, despite the fact that it could perceive a virtual world, make inferences about that world, and even justify its knowledge to a limited degree. Such extra knowledge would seem to be quite unnecessary to a rational system, and would only complicate the processes of awareness and inference. (Hofstadter 1979 discusses this point at length.)

Instead, it seems very likely that such a system would have just the same kind of attitude towards its perceptual contents as we do toward ours, with its knowledge of them being direct and unmediated, at least as far as the system is concerned. When we ask how it knows that it sees the red tricycle, an efficiently designed system would say "What do you mean, how do I know? I just *see* it!". When we ask how it knows that the tricycle is red, it would say the same sort of thing that we do: "It just looks red." If such a system were reflective, it might start wondering about how it is that things look red, and about why it is that red *just is* a particular way, and blue another. From the system's point of view it is just a brute fact that red looks one way, and blue another (of course from our vantage point we know that this is just because red throws the system into one state, and blue throws it into another; but this does not help from the machine's point of view).

As it reflected, it might start to wonder about the very fact that it seems to have some access to what it is thinking, and that it has a sense of self. In short, a reflective machine that was designed to have direct access to the contents of its perception and thought might very soon start wondering about the mysteries of consciousness. "Why is it that heat *feels* this way?"; "Why am I *me*, and not someone else?"; "I know my processes are just electronic circuits, but it certainly seems to me that there's something extra." (This idea is developed in a very compelling way by Hofstadter 1985.)

Of course, the speculation I have engaged in here is not to be taken too seriously, but it does bring out the great *naturalness* of the fact that we judge and claim that we are conscious, given a reasonable design. It would be a strange kind of cognitive system that had no idea what we were talking about when we asked what it was like to be it. The fact that we think and talk about consciousness may be a consequence of very natural features of our design, just as it is with these systems. And certainly, in the explanation of why these systems think and talk as they do, we will never need to invoke full-fledged *consciousness*. Perhaps these systems are really conscious and perhaps they are not, but the explanation works independent of this fact. Any explanation of how these systems function can be given solely in computational terms. In such a case it is obvious that there is no room for a ghost in the machine to play an explanatory role.

All this means that we can lay out a revised version of a principle put forward in Chapter 1.

The Surprise Principle (revised version): Consciousness is surprising. Claims about consciousness are not.

Consciousness is a feature of the world that we would not predict from the physical facts, as we have seen. If it were not for the fact that conscious experience is a brute fact presented to us directly, there would be no reason to postulate its existence. By contrast, the things we say about consciousness are a garden-variety cognitive phenomenon. Somebody who knew enough about cognitive structure would immediately be able to predict the likelihood of utterances such as "I *feel* conscious, in a way that no physical object could be", or even Descartes' "Cogito ergo sum". In principle, some reductive explanation in terms of internal processes should render claims about consciousness no more deeply surprising than any other aspect of behavior. I have gestured toward such an explanation above, and will consider the matter in more detail in later work.

We will see later that the details of an appropriate explanation can be very useful in getting a theory of consciousness off the ground. The relationship between an explanation of phenomenal judgments and an explanation of consciousness is a subtle one, however. Before proceeding, I will consider a more unsubtle response to the situation we are placed in.

5.5 Is explaining the judgments enough?

At this point a natural thought has probably occurred to many readers, especially those of a reductionist bent: If one has explained why we *say* we are consciousness, and why we *judge* that we are conscious, haven't we explained all that there is to be explained? Why not simply give up on the quest for a theory of consciousness, declaring consciousness itself a chimera? Even better, why not declare one's theory of why we judge that we are conscious to be a theory of consciousness in its own right? It might well be suggested that a theory of our judgments is all the theory of consciousness that we need.

This position gets some support from considerations about judgments in other domains. It might be thought that the widespread belief in gods, found in all sorts of diverse cultures, provides an excellent reason to believe that gods exist. But there is an alternative explanation of this widespread belief in terms of social and psychological forces. One can appeal to people's psychological insecurity in the face of the cosmos, to the need for a common outlet for spiritual or emotional expression, and to the intrinsically self-propagating nature of certain idea-systems to explain why it is all but inevitable that religious beliefs should be widespread, given our nature and circumstances. One can even point to the existence of certain highly plausible but faulty arguments for the existence of a god, such as the argument from design and the cosmological arguments. Although these arguments are faulty, they are not *obviously* faulty (in particular, the argument from design could reasonably have been seen as compelling before the time of Darwin), and it is not hard to see why they should generally contribute toward the naturalness of religious belief.

The observation that widespread religious belief can be explained in this way, without appeal to the existence of any gods, is generally taken to provide further evidence that no gods in fact exist. As it turns out, the atheistic hypothesis not only can explain the complex structure of nature as well as the theistic hypothesis; it can even explain why the theistic hypothesis is so popular! This is a tremendous way of cutting the ground from underneath an opposing view, and in the case of religious belief, the argument seems quite compelling.

But the analogy with consciousness fails. Explaining our judgments about consciousness does not come close to removing the mysteries of consciousness. Why? Because consciousness is itself an *explanandum*. The existence of God was hypothesized largely in order to explain all sorts of evident facts about the world, such as its orderliness and its apparent design. When it turns out that an alternate hypothesis can explain the evidence just as well, then there is no need for the hypothesis of God. There is no separate phenomenon *God* that we can point to and say: *that* needs explaining. At best, there is indirect evidence. (I leave aside so-called "religious experiences" here. One can fairly doubt that such experiences are really a direct experience of God. It seems tenable to suppose that these are merely experiences of deep spirituality and awe, and that the subjects in question unconsciously infer that God must be the source of such glory.) Similarly, the existence of UFOs is often postulated to explain strange events in the sky, markings in the ground, disappearances of ships and planes in the Bermuda triangle, and so on. If it turns out that this evidence can be explained without postulating the existence of UFOs, then our reason for believing in UFOs disappears.

Consciousness is not like this. As we have seen, it is not an explanatory construct, postulated to help explain behavior or events in the world. Rather, consciousness is a brute explanandum; it is a phenomenon in its own right that is in need of explanation. It therefore does not matter if it turns out that consciousness is not required to do any work in explaining other phenomena; our evidence for consciousness never lay with these other phenomena in the first place.

Even if our judgments about consciousness are explained reductively, *all* this shows is that our judgments can be explained reductively. The mind–body problem is not that of explaining our judgments about consciousness—if it were, it would be a relatively trivial problem, and it would be hard to see why everybody had been so worried about it. The key explanandum in the mind–body problem is not the existence of judgments about consciousness, but consciousness itself. If the judgments can be explained without explaining consciousness, then that is interesting and perhaps surprising, but it certainly does not remove the mind–body problem.

To take the line that explaining our judgments about consciousness is enough (just as explaining our judgments about God is enough) is most naturally understood as an eliminativist position about consciousness (as one analogously takes an eliminativist position about God). As such it suffers from all the problems that eliminativism naturally faces: in particular, it denies the evidence of our own experience. To deny the existence of conscious experience is the sort of thing that can only be done by a philosopher or by someone else tying themselves in intellectual knots. Our experiences of red do not go away upon making such a denial. It is still like something to be us, and that is still something that needs explanation. To throw out consciousness itself as a result of the paradox of phenomenal judgment would be to throw out the baby with the bathwater.

There is a certain intellectual appeal to the position that explaining phenomenal judgments is enough. It has the feel of a bold stroke that cleanly dissolves all the problems, leaving our confusion lying on the ground in front of us exposed for all to see. Yet it is the kind of "solution" that is satisfying only for about half a minute. When we stop to reflect, we realize that all we have done is to explain certain aspects of our behavior. We have explained why we talk in certain ways, and why we are disposed to do so, but we have not remotely come to grips with the central problem, namely conscious experience itself. When half a minute is up, we find ourselves looking at a red rose, inhaling its fragrance, and wondering: "Why do I experience it like *this*?". And we realize that our explanation has nothing to say about the matter.

If this position is not taken as a kind of eliminativism, it can perhaps be taken as a kind of functionalist position, on which the notion of consciousness is construed as 'the thing responsible for our judgments about consciousness'. But like any other functional definition of consciousness, this seems to be quite inadequate as an account of our concept. Whether or not consciousness is *in fact* responsible for our judgments about consciousness, this does not seem to be a conceptual truth. This is brought out by the fact that while the notion of explaining our judgments without explaining consciousness is quite counterintuitive, it is certainly not *incoherent*. It is at least a logical possibility that one could explain our judgments without explaining consciousness; and that is enough to show that this construal of consciousness is a false one.

There are other variations on this line of argument. For instance, one could argue that there is a purely reductive explanation of why I think that consciousness cannot be reductively explained, or of why I think consciousness is not logically supervenient, or of why I think it cannot be functionally defined. It might even explain why I think conscious experience is an explanandum. This might be thought to undercut my arguments in earlier sections entirely, opening the way for a reductive view of consciousness. But again this view can be satisfying only as a kind of intellectual cut-and-thrust. At the end of the day, the fact is that consciousness is still there, it still needs to be explained, and an explanation of behavior or of some causal role is simply explaining the wrong thing. This might seem to be mule-headed stubbornness, but it is grounded in a simple principle: our theories must explain what cries out for explanation. We have seen earlier that there are very good reasons to believe that no reductive account can explain consciousness. Those who dispute this had better find something wrong with the arguments; an appeal to the psychology of the arguer is not enough.

Anybody who wants to take the line that explaining phenomenal judgments is enough had best go back to previous chapters, then, and find something wrong with the arguments there. If the arguments there succeeded, then this argument cannot count against them. If on the other hand the earlier arguments failed, then this argument is unnecessary. Either way, this argument does not advance the state of play. The puzzle of consciousness cannot be removed by such simple means.

One advocate of the position that our judgments about consciousness are all we need to explain is Daniel Dennett. In Dennett (1979) he writes:

> I am left defending the view that such judgments *exhaust* our immediate consciousness, that our individual stream of consciousness consists of nothing but such propositional episodes, or better: that such streams of consciousness, composed exclusively of such propositional episodes, are the reality that inspires the variety of misdescriptions that pass for theories of consciousness, both homegrown and academic. (p. 95)

and

> My view, put bluntly, is that there is no phenomenological manifold in any such relation to our reports. There are the public reports we issue, and then there are the episodes of our propositional awareness, our judgments, and then there is—so far as introspection is concerned—darkness. (p. 95)

To this, all I can say is that Dennett's introspection is very different from mine. When I introspect, I find sensations, experiences of pain and emotion, and all sorts of other accoutrements that, although *accompanied* by judgments, are not *only* judgments—unless one *redefines* the notion of judgment, or of "episodes of our propositional awareness", to include such experiences. If so, then Dennett's position is reasonable, but there is no longer any reason to suppose that our judgments can be reductively explained. If judgments are instead construed as functionally individuated states such as dispositions to report, as I think Dennett intends, then his thesis becomes vastly implausible, simply consisting in a denial of the data that a theory of consciousness must explain.

What might be going on when someone claims that introspection reveals only judgments? Perhaps Dennett is a zombie. Perhaps he means something unusual by 'judgment', as above. Most likely, however, he has taken something else for introspection: what we might call *extrospection*, the process of observing one's own cognitive mechanisms "from the outside", as it were, and reflecting on what is going on. Observing one's *mechanisms*, it is easy to come to the conclusion that it is judgments that are doing all the work. All that is going on in the relevant cognitive processes is a lot of categorization, distinction, and reaction. The processes involved with my perception of a yellow object can plausibly be fully explained in terms of certain retinal sensitivities, transformations into internal representations, and categorization and labeling of these representations. But this does not explain the contents of introspection; it explains only the *processes* involved. Extrospection is not introspection, although it is easy to see how a philosopher inclined to speculate on his own internal mechanisms could take one for the other. Conscious experience remains untouched by this explanatory method.

In general, when one *starts* from phenomenal judgments as the explananda of one's theory of consciousness, one will inevitably be led to a reductive view.[4] But the explananda are not the judgments, but conscious experiences themselves. This sort of approach "solves" the problem only by a misdirection of attention.

Considerations about phenomenal judgments certainly provide the greatest pressures and even the best arguments in favor of a reductive view of consciousness. On the face of it, it is hard to see how we could know about consciousness if it does not play an explanatory role in our judgments. But considerations about consciousness itself provide compelling arguments that no reductive explanation can succeed. The question then is: How can we reconcile the failure of reductive explanation with our knowledge of consciousness? This is a tricky matter and not entirely obvious, although it turns out not to be as hard as one might think. But we cannot put the cart before the horse by ignoring consciousness entirely, and worrying only about the judgments.

There is certainly an important insight in the idea that our judgments about consciousness may provide the key to a theory of consciousness. But it will not be as simple as this. Instead of imagining that we can use these judgments to push consciousness off the stage entirely, we must give them a central role in a theory that takes consciousness seriously. With the initial defensive moves out of the way, I will begin this project in the next chapter, and continue it in the remainder of this work.

5.6 Appendix 1: Further evidence for explanatory irrelevance

The argument that I have given for the explanatory irrelevance of consciousness is an indirect one. The conclusion is forced on us by independent evidence for two different theses: the causal closure of physics, and the failure of consciousness to logically supervene. Despite this intellectual basis, some people will find the conclusion counterintuitive and very difficult to accept.

In this appendix, I will discuss some independent considerations that tend to

[4]Other advocates of a reductive view of consciousness have been impressed by the fact that our phenomenal judgments can be explained in physical or functional terms. For example, Foss (1989) responds to Jackson's knowledge argument by noting that Mary, the colorblind neuroscientist, could know everything that a subject with color vision would *say* about various colors, and even everything that a subject *might* say. But of course this falls far short of knowing all there is to know.

support the thesis. These will reinforce its plausibility, and perhaps help counteract the sense that something must have gone wrong in the arguments that got us this far. These include considerations about the ineffability of conscious experience, contemporary theories of perception, and evidence from experimental psychology.

The ineffability of conscious experience

We have seen already that conscious experiences are remarkably hard to describe. Despite the seemingly rich nature of one's experience of red, for instance, there is very little one can say about that nature. The quality of redness seems to elude description entirely.

Conscious experience is not utterly indescribable, of course. First, we can describe experiences in terms of their *relational* properties. In particular, we frequently catalog experiences based on what they are typically caused by. We can speak of a fruity taste, or of a rotten-egg smell. Even a description of an experience as "red" effectively has its reference fixed by its typical cases. Another sort of relational description tags an experience by the psychological state with which it is usually associated. We can talk of an angry feeling, or of hunger pangs. But none of these descriptions say anything about the intrinsic nature of the experience, as opposed to its extrinsic relations. They do not describe what the experience *feels like*. At best, these descriptions can communicate the intrinsic nature to someone whose experiences have the same relational properties: we say "angry feeling", meaning "the experience I have when I am angry", and they say "Ah, *that* feeling".

We can also describe complex experiences in terms of their *structural properties*. I can describe the structure of my visual field in terms of various geometric relations, or I can describe the temporal structure of a musical experience. We can also speak of the intensity of an experience, which can be regarded as another structural property. Further, we can describe relations of similarity and difference between different experiences, as with colors, which perhaps represent a kind of implicit structure,

But when it comes to the brute nature of a simple experience, such as the quality of redness, there is nothing we can say. We are at a loss for words. There certainly *is* a rich, qualitative way in which red is quite different from green, and different again from yellow. But there is almost nothing we can say about this, other than the fact that it *is* different.

It is a remarkable property of consciousness that for something with such a central place in our life, there is so little we can say about it. An examination of the taxonomy I put forward in Chapter 1, for instance, reveals very little in the way of descriptions of experiences, over and above descriptions of structure and relational properties. For the most part, in asking why an experience has the particular quality that it does, all I could do was ask: why does it feel like *that*?

Insofar as we can describe features of consciousness at all, it is only insofar as they are mirrored by features of awareness; structural or relation features, for instance. The most striking features of experience (the quality of redness, the unique timbre of middle C) cannot be described, and these are precisely the features that are not mirrored in awareness.

All can be seen as supporting the explanatory irrelevance of consciousness. The fact that the most central and interesting features of experience cannot make it into our descriptions supports the notion that these features play no role in later processing. Even when features of experience *are* mirrored in our descriptions, as with structural properties, this effect can be explained in terms of the structure of awareness. Either way, what is special about consciousness seems to play no role in our descriptions.

Another way to note the ineffability of experiences is to imagine an inverted spectrum case, in which you have red experiences where I have green experiences, but our behavior is more or less the same (perhaps your optic nerve has been rewired since birth). In such cases, there would be a huge difference in the nature of our color experience on looking at grass, but it would not be manifested in anything we say. By hypothesis we would say the same sorts of things, including "That looks nice", "An appealing bright patch over there", "Isn't green an interesting color?", and so on. The difference between red and green experiences is simply not the sort of thing that makes it into verbal reports. On the basis of the things we say, we might never know the difference in our experience. (See Taylor 1966 for more on the implications of inverted spectra for the incommunicability of experience.)

Inverted spectrum thought-experiments can bring out the irrelevance of the redness of a red experience to later processing as well as to verbal reports. If all our red experiences were replaced by green experiences, it is not obvious that anything in later processing would thereby change. One can see this intellectually by noting the conceptual possibility of an inverted replica of me, but the possibility is also plausible on more direct grounds. There seems to be nothing in the role that red experiences plays in my system that is especially indicative of its red nature, as opposed to a green nature. Certainly there are judgments of "that's red", and various associations, but it is plausible that if I had always had green experiences in the place of red, those verbal judgments and associations would be precisely the same.

One can even find the same phenomenon in introspection. Of course in our introspection of the experience itself, the red nature is extremely salient. But when we examine the *thoughts* that we have about the experience, there is little that seems to depend on that nature. I can think "I really appreciate experiences like *that*", and I can think "*that* is the kind of thing that could not be physical", and so on, but all these are thoughts that might be quite unchanged if a green experience were in the red one's place.

Indeed, when I am appreciating the unique appeal of a particular experience, such as a sensation of green, I sometimes find myself thinking: the *form* of these thoughts is exactly the form of the thoughts that I have about red. Only the pointer is different. The thoughts are always: "*that*'s beautiful", "what an interesting intrinsic quality *that* has". The quality certainly permeates my introspection, but *only* in the experience. It does not seem to find its way into the judgment. Any other quality with the same "structural" properties (intensity, similarities and differences with other experiences, causal relations) might have produced similar judgments.

If someone questions the hypothesis of the explanatory irrelevance of experience, these phenomena of ineffability seem to bear it out. There is very little about the judgments that we form that seems to *require* any special quality to play a role. Only structural properties are required, and those are already present in the structure of awareness. At best the redness of the experience is reflected in judgments like "feels like *that*", but the very fact that such an ostensive pointer is required suggests that the properties of the experience have not made a difference to the judgment. Certainly there is a difference to *me*, as I am more than a bundle of judgments, but it is arguable that the only difference lies in my direct experience. It is remarkably hard to find a role for the specific experience to play.

All this is very speculative, but it might seem to bear out the notion that *some* kind of epiphenomenalism is the case. Any role played by the experience is at best a very thin one.

The role of experience in perception

Further support for the supposition of explanatory irrelevance is suggested by contemporary theories of the role that experience plays in perception. Traditional accounts of perception allowed conscious experience a central role. Often in such accounts, sensations or qualia were the fundamental units on which all units were founded. In "sense-data" theories, for instance, our perception is first and foremost of various sensory patches, and it is only indirectly from these that we "infer" the world.

Contemporary theories of perception have almost uniformly rejected this view (although see Jackson 1977). There does not seem to be any room for special internal "objects" of perception. Instead we perceive the external world more or less directly, certainly mediated by a causal process, but without any special stopping-point at the sensation. To argue that all perception proceeds via perception of sensation seems to involve at worst a regress, and at best an unparsimonious two step account that is quite unnecessary.

These contemporary theories are quite compatible with the explanatory irrelevance of experience. On sense-data theories, there was room for a straightforward cognitive role for sensations. Objects in the environment caused sensations, and sensations were the basis for our later judgments. By cutting sense-data out of the loop, it seems that the most natural cognitive role that one might suggest for experiences has been eliminated. It is true that this does not rule out every role that those experiences might play, but it does add plausibility to the hypothesis that there is no substantial role for them.

Dennett (1991) attacks "qualophiles" as if they were committed to the thesis that qualia are the grounds on which perceptual judgments are based. But we have seen that the friend of qualia is committed to no such thing. In fact, on the most plausible position, qualia do not play any special role in the formation of those judgments, but are simply a concomitant of those judgments. The story about perceptual awareness can be told in terms of first-order judgments alone. By structural coherence, qualia will have a parallel structure, but we do not need them to enter the picture. Qualia are not posited in order to explain perception; as we have seen, they are brute explananda. These attacks therefore miss their mark.

The last century of philosophy has seen a series of attacks on the "given": the notion that perception is grounded in sense-data, and the notion that our knowledge of the external world is an inference from "given" conscious experience. These arguments are sometimes seen as attacks on the notion of experience itself, but I interpret them differently. Insofar as these are successful, they are generally *extrinsic* attacks; they say that positing direct experience is not *useful* for the explanation of perception or of knowledge. Such arguments rarely focus on the question of whether there is conscious experience; they focus only on the role that it plays in these theories. The arguments are therefore compatible with my position, and can even be seen to support it. Given the explanatory irrelevance of consciousness, there will be no mileage in basing a theory of external perception on conscious experience, or in doing the same for knowledge. We can therefore see these arguments as providing further support for the explanatory irrelevance of consciousness.

Evidence from experimental psychology

Many cognitive psychologists have noted that their models do not postulate any role for consciousness. Any function that we might think would be best explained by consciousness is inevitably explained by some information-processing mechanism that seems quite compatible with the absence of consciousness. Within psychology and artificial intelligence, it is therefore common to regard consciousness as an "epiphenomenon", although the term is often used loosely without any special ontological commitment.

Of course, these observation are essentially a recapitulation of the argument for the explanatory irrelevance of consciousness given already. The relevant functional processes are causally closed, so explanations of behavior can be given solely in terms of such processes. The observation that this gives no role for consciousness is merely recaps the observation that consciousness is not itself logically supervenient on these processes. So it is no surprise that such a role should be difficult to find, but it is interesting nevertheless.

More specific arguments for the explanatory irrelevance of consciousness in cognitive psychology have been put forward by Velmans (1991). These arguments appeal to a host of experimental data to support the conclusion that consciousness has no function in the cognitive system. I am skeptical whether the empirical evidence plays any special role in the argument over and above the general argument we have already given, but it is still interesting to examine Velmans' argument briefly.

Velmans uses experimental data to make two points. The first is that for almost any function that one might expect to be a function of consciousness, one sometimes finds the function being performed in the absence of consciousness. He presents empirical evidence that the following functions can be performed in the absence of conscious awareness of the relevant information: (a) analysis of input stimuli; (b) selection from competing stimuli; (c) learning (d) encoding of information into memory; (e) the control of action; (f) planning; and (g) creativity. The second part of his argument is that even where the data suggest that certain functions almost invariably go along with consciousness, what is really responsible for the function is *focal-attentive processing.* He argues that consciousness is merely a result of focal-attentive processing, rather than being identical with it, so that explanations in terms of focal-attentive processing remove the need to make any explanatory need to consciousness.

We should be given pause by noting that if this sort of empirical evidence establishes the irrelevance of consciousness, it should equally well establish the irrelevance of its functional correlate of *awareness*; after all, the two are perfectly correlated, so performance of a function without consciousness implies performance without awareness. But this is absurd: awareness is functionally defined, after all, and *obviously* has some causal role, not least in the production of verbal reports. It is interesting that Velmans never considers the causal relation between consciousness and verbal reports; the relevant experimental paradigms all *presuppose* that reports are evidence for consciousness, and therefore can never provide direct evidence against such a causal connection. On the face of it one would also imagine that awareness is also responsible for the production of integrated rational action of the kind that goes along with considered thought. Velmans also presents no empirical evidence against this conclusion.

It seems that what is really doing the work for Velmans is the philosophical assumption that consciousness is *independent* of such functionally-defined notions. His notion of focal-attentive processing is very nearly my notion of awareness, with a dollop of attention thrown in so that objects of "background awareness" are excluded.

In arguing that explanations in terms of focal-attentive processing exclude an explanatory role for consciousness, he presupposes that consciousness is not logically supervenient or otherwise conceptually related to focal-attentive processing; if it were, his argument could not establish the desired conclusion.

It follows that empirical evidence is not really central to Velmans' argument, despite appearances. Given any functionally defined process, we can say: that is not *consciousness*. Explanations in terms of such processes will therefore never reveal a role for consciousness. This part of the argument is all but *a priori*. Responses to this sort of argument must inevitably fall onto philosophical rather than empirical ground.

It is nevertheless interesting to see the large number of functions that can be performed without consciousness. This perhaps helps reinforce the conclusion that consciousness did not come into existence primarily in order to play some causal role. Although I am skeptical about whether the empirical evidence is playing a central role in the argument, this sort of consideration can nevertheless help buttress the conclusion simply by allowing our abstract intuitions about the concept of consciousness to be fleshed out against a more specific background. If *a priori* conclusions are recapitulated in arguments from empirical evidence, the argument is at least made more robust by being played out on a wider playing-field.

There seems to be a convergence of support, then, for the thesis that experiences are explanatorily irrelevant to cognition. Any role they might play is thin at best. In fact the only *prima facie* role that seems to be left is a role in enabling our judgments about those very experiences; a role which is certainly incidental to most of our cognitive abilities. In the above, we have seen that even this effect might be explained without invoking a special role for experiences themselves in the explanation. It therefore seems that explanatory irrelevance of experience may not only be the right intellectual conclusion, forced on us by the failure of logical supervenience and the causal closure of physics. It may also be the conclusion that makes the most sense.

5.7 Appendix 2: The content of phenomenal judgments

It might be thought that there are problems with the *content* of second-order and third-order phenomenal judgments. In particular, we want to say that our phenomenal notions *refer* to our conscious experiences, so that our phenomenal judgments can be made true or false by facts about those conscious experiences. But reference is often understood causally, so that a notion refers to an entity when there is an appropriate causal relation between them, and we have seen that any causal role that conscious experience plays is elusive at best. Might there then be a problem with our judgments being *about* conscious experience?

I do not think there is a problem. Although causal relations often ground the reference relation, this need not always be the case. For instance, we can refer to the largest star in the universe, even though we may have no causal connection with that star. What is important is that our concepts possess (prior) *intensions*, and that our judgments possess *truth-conditions*. If the concepts in question have appropriate intensions, reference will be determined and the judgments will have truth-value.

We saw in chapter 2 that the conditions that go into specifying the intension of a notion will often centrally involve a causal relation. For instance, the prior intension of water may require that the referent bear an appropriate causal relation to us. But intensions need not have this property, as the example of 'the largest star in the universe' shows. All that matters is that an intension is sufficiently determinate. If so, reference will be fixed. Similarly, all that matters for judgments is that they have truth-conditions sufficient to determine truth or falsity in the actual world. The primary determinants of content, then, are the prior intensions that fix reference, and the relevant truth-conditions. From these the story about reference and truth will be understood.

The presence or absence of consciousness is irrelevant to the determination of this intension. The prior intension associated with 'water' would be the same even if water did not exist, although the referent and the posterior intension would be quite different. Indeed, it seems plausible to say that phenomenal judgments made by a zombie have precisely the same (prior) truth-conditions as mine. There is a clear sense in which when my zombie twin claims that he is conscious, he is claiming for himself just what I am claiming for myself; he is just wrong about it. We want an account of the truth-conditions of phenomenal judgments so that his come out to be false, whereas mine come out to be mostly true. It follows that the presence or absence of consciousness does not contribute to the (prior) truth-*conditions* of these judgments, although it may contribute to their *truth*. Explanatory irrelevance therefore poses no problem for content.

An account of the truth-conditions of phenomenal judgments is difficult to give explicitly, but its basic form is straightforward. The truth-conditions are as one would expect. A belief "I am conscious" is true if and only if the bearer is conscious. There is not much more one can say about it than this; the notion of consciousness seems to be irreducible, as we have seen.

There is also the matter of posterior truth-conditions. We saw in Chapter 2 that there are two sorts of truth-conditions associated with any statement. The prior truth-conditions (determined by the prior intensions involved) determine whether the statement is true of the actual world, depending on how that world turns out. The posterior truth-conditions determine its truth across possible worlds, given that the actual world is fixed. It is the prior intensions and truth-conditions with which I have been concerned above, as it is these that fix reference and truth. In any case, we saw in Chapter 4 that the prior and posterior intensions of 'consciousness' are identical. It follows that the two sets of truth-conditions in this case are the same.

The truth-conditions of a belief such as "I am having a red sensation" are more complex, as the notion of "red sensation" can be further analyzed. We saw in Chapter 1 that the prior intension of 'red sensation' can be roughly understood as "the kind of phenomenal property caused (in me) by such-and-such objects in the environment", where the objects in question are roses, fire trucks, and whatever else the subject used in acquiring the concept. An analysis like this determines the prior truth-conditions for a judgment about red sensations. (Note that the truth-conditions are not entirely functional, due to the occurrence of the notion of "phenomenal property" within.)

The posterior intensions of such judgments can differ from the prior intensions in these cases. If the relevant class of objects causes one kind of sensation in one person, and another in someone else—as might happen if inverted spectrum cases are not just possible but actual—then it will turn out that 'red sensation' picks out different referents for each of them, and that their posterior intensions will differ accordingly. It follows that the posterior truth-conditions are dependent on the intrinsic nature of the sensations in question. As for posterior truth-conditions of such judgments in a zombie, these may be somewhat loose and indeterminate, due to the failure of the prior intension to pick out an actual-world class. Perhaps these can be understood by analogy to the posterior intension of the term 'unicorn'.

To see how this would apply to inverted spectrum cases, imagine that when looking at grass you have the kind of color experience that I have when looking at blood, and vice versa. You say "grass gives me green experiences" and you say it truthfully, despite the fact that if *I* say "grass gives you green experiences", I am speaking falsely. The term 'green experience' has a different referent and therefore a different posterior intension for the two of us, due to the indexical element in the prior intension. For you, a "green experience" is the kind of experience caused by green things in *you*; for me, a "green experience" is the kind caused by green things in *me*. Due to this indexical element, when you say "grass gives me green experiences" and I say "grass gives you red experiences", we can both speak truthfully[5] (contra Taylor 1966). (Compare: you say "I am hungry" and I say "I am not hungry"; or I say "water is H_2O" and my Twin Earth twin says "water is XYZ".)

[5]This relativism does not hold for *external* color terms, such as 'green' as a property of *objects* rather than experiences. At a first approximation, an object is green if for most of us, it tends to gives rise to the same sort of experiences in standard conditions as grass, trees, and so on do. The posterior intension of 'green object' is therefore *public* in a way that the posterior intension of 'green experience' is not'; it is defined relative to a community rather than an individual. Even someone with an inverted spectrum would use "green object" with the same posterior intension as me. If you

In any case, none of this requires a causal theory of reference. In each case, at least the prior intension is determined fully by our cognitive structure, and is shared between ourselves and zombies. The prior truth-conditions of these judgments are also determined fully by cognitive structure, then, and can be explicated roughly as above. The question of just *how* our cognitive structure determines intensions is of course a very difficult one that is at the crux of the theory of meaning in the philosophy of mind. So far, we do not have any good theories of this. I will not even touch this project, being happy to simply accept that intensions are somehow determined for judgments about consciousness as they for judgments about plants or rocks.

One clear way in which the contents of phenomenal judgments differ between my zombie twin and myself is in *reference*. My judgments about "my conscious experiences" refer, whereas his do not. This can be understood compatibly with the notion that our judgments share the same prior truth-conditions. It simply turns out that in the (centered) world of the zombie nothing satisfies the relevant intension, so that his notion has no reference.

It is possible that there is some other way in which the contents of phenomenal judgments differ between zombies and ourselves. If so, this will be in an aspect of content that lies beyond truth-conditions. It is plausible that there are such aspects, but the matter is very poorly understood. I will not pursue it here.

---

say "that object is green" and I say "that object is red", we are not both speaking truly.

We might even make the analysis purely functional by saying that an object is green if normal observers will tend to *judge* it to look the same color as grass under normal circumstances (or: will judge it to produce the same kind of experience). This has the analysis of allowing zombies to speak truly about green objects, as perhaps seems reasonable. After all, it seems that intersubjective similarity in *judgments* rather than similarity in experience is all that is required to get the reference of external color terms going.

We might allow for a small amount of relativism, so that if you and I have slightly different boundaries in our color discrimination, there might be a sense in which an object could be blue (for me) and green (for you). Note that this is dependent on a *functional* difference between us, though, rather than the pure phenomenal difference that inverted spectra provide. The very existence of color terms reflects the fact that functionally, our color boundaries are fairly similar, so that these indeterminacies will be slight. The same indeterminacy in reference could not hold for a case where someone had very unusual color boundaries; see below.

A fully relativistic view will not work: if someone, even someone with very different experiences, says "grass is red", they are speaking falsely. As the terms are used in our linguistic community, it is an objective fact that grass is green. The term "green experience" used by someone with an inverted spectrum refers to red experiences, but "green object" still refers to green objects. If a person were "rewired" in a way so that their color boundaries were different to mine, so that we agree on "grass is green" but they say "frogs are red", they are simply *wrong* in making the second claim, even though they may be right in saying that frogs produce red experiences in them, or even that frogs look red to them.

Chapter 6

On the Coherence between Consciousness and Cognition

6.1 Principles of coherence

Perhaps the most promising approach to developing a theory of consciousness focuses on the remarkable *coherence* between conscious experience and cognitive structure. The phenomenology of mind, typified by our conscious experiences, and the psychology of mind, typified by phenomenal judgments, do not float free of each other in practice but are systematically related. The many lawful relations between phenomenology and psychology can provide much of what we need to get a theory of consciousness off the ground.

This coherence manifests itself in many forms. Perhaps the simplest of these is what I will call the *principle of reliability*: our second-order judgments about consciousness are usually *right*. When I judge that I am having a blue sensation, I am usually having a blue sensation. When I think I have just experienced a pain, I have usually just experienced a pain.

This is not to say that we are absolutely incorrigible in our judgments about consciousness. If I am distracted, sometimes I might believe that I have just experienced pain when in fact all I experienced was loud noise; on the flip side, in circumstances where I am focusing elsewhere I can believe I have experienced *no* pain when a pain has just occurred. Reliability becomes even more questionable in cases of mental illness, or of various neurophysiological pathologies. I can also *mislabel* experiences, just as one can mislabel anything, if I do not have a good grasp of the relevant category. I might mislabel a crimson experience maroon if I were not good with color names, for instance. Nevertheless, for normally functioning humans in normal circumstances, who grasp the relevant categories appropriately, phenomenal judgments are generally correct. This correctness happens far too often to be an accident. Something systematic must be going on.

In the reverse direction, it seems that when we have a conscious experience, we generally have the capacity to notice it, remember it, comment on it, and so on. This I will call the *principle of detection*. Of course many or most of our experiences pass by without our paying much attention to them, but it would be an odd sort of experience that was *in principle* unnoticeable by us. That experiences are generally noticeable seems to be a basic property of consciousness.

Again, this claim is not absolute. There are arguably experiences that flicker by so fast that they cannot be remembered—we may know that something has happened, but we do not have time to focus our attention on just what, and the details are lost to memory. (Dennett 1991 provides an interesting catalog of cases that can be interpreted this way.) Even in these cases, it is arguable that there is some momentary noticing, although not traces are left in short-term memory. In any case, it seems that we are capable of noticing any experience that lasts a reasonable length of time if we direct our attention appropriately.

Some might question the grounds for this claim. After all, I do not have access to the experiences of others. But as we saw earlier, the facts about one's own case can take us a long way if we supplement them with principles of plausibility, homogeneity, and simplicity.

My evidence for the principles of reliability and detection is grounded in my own case. For the principle of detection: I do not have any reason to believe I have conscious experiences that I cannot notice, and wielding Ockham's razor I can discard the possibility. The job of a theory of consciousness is only to explain the phenomena, and unnoticed experiences are not among the phenomena. In a similar way, it squares with my introspection that my judgments about sensations are generally correct, and that the principle of reliability holds. Someone might argue that there is something circular about this—after all, *I* am judging *my* judgments correct—but it is hard to see how things could be otherwise. In any case, if my judgments about my conscious experiences were systematically *wrong*, it is hard to see how a theory of consciousness could get started. It follows that not only are the principles of reliability and detection borne out by introspection, but they also play the role of basic methodological requirements on obtaining a theory of consciousness.

It is true that my introspection gives evidence only about my own case *now*, but if we invoke a principle of homogeneity saying that the present is not especially different from the past, it seems likely that these principles have held in the past and will continue to hold in the future. Furthermore, I certainly *remember* them holding in the past; while these memories may be suspect, the general reliability of memory not only is vastly plausible, but also has the status of a methodological principle in these investigations for reasons similar to those above.

The extension of these results from my own case to that of others once again proceeds on grounds of homogeneity and plausibility. There is no reason to believe that I, David Chalmers, am *special*; why should the laws of consciousness be any different with me than with anyone else? The universe would be a strange, arbitrary place if they were. On the grounds that consciousness in others works like consciousness in me, it is reasonable to suppose that these systematic relations hold in others too.

Apart from these philosophical reasons, it simply squares with common sense that for the most part when people say they are having sensations they are having sensations, when people have experiences they can notice them, and so on. Of course, common sense should not be overrated in these matters. At various points in this work, I have arrived and will arrive at conclusions that arguably contradict common sense. Each time, this is because there are compelling grounds to do so, generally in the form of positive arguments that some counterintuitive state of affairs is the way things are. But common sense should not be underrated, either. *Other things being equal*, we should come down on the side of common sense.

Anyone questioning the principles above is left in a position like that of the solipsist about consciousness, or one who denies the existence of the external world. These are consistent positions that cannot be disproved, but that are at odds with everything else we believe, and that entirely lack positive motivation. Given the lack of independent reasons to believe in these positions, over and above their consistency with the evidence, we have no trouble counting them out as serious contenders. The same goes for the coherence principles above. That these principles hold is simply the most natural way for the world to be that is compatible with the evidence, even though we cannot logically exclude the more bizarre alternatives.

6.2 Coherence between consciousness and awareness

The most fundamental coherence principle between consciousness and cognition does not involve second-order phenomenal judgments. Rather, it centers on the relationship between consciousness and our first-order judgments. The principles with which we will deal here concern the coherence between consciousness and *awareness*. Recall that awareness is the psychological correlate of consciousness, roughly explicable as a state wherein we have direct access to some information, and wherein this information is available for the deliberate control of behavior and for verbal report. The contents of awareness correspond precisely to the contents of our first-order phenomenal judgments, the contentful states that are not about consciousness but that are parallel to it.

Wherever there is consciousness, there is awareness. My visual experience of a red book upon my table is accompanied by a functional *perception* of the book. Optical stimulation is processed and transformed, and eventually a judgment is made that there is a book of such-and-such shape and such-and-such color on the table, with this information available in the control of behavior. The same goes for the details in what is experienced. Each detail is cognitively represented. To see that each detail must be so represented, simply observe that I am able to comment on those details and to direct my behavior in ways that depend on them (for instance, I can point to appropriate parts of the book). Such systematic availability of information implies the existence of an internal state carrying that content.

The same goes for any sensory experience. What is experienced in audition is represented in our auditory system, in such a way that later processes have access to it in the control of behavior; in particular, the contents are available for verbal report. In principle, somebody who knew nothing about consciousness might examine our cognitive processes and ascertain these contents of awareness by observing the role that information plays in directing later processes. In the same sort of way we can even handle hallucinations and other cases of sensations without a real object being sensed. Although there is no real object for the contents of perception to concern, there is still representation in our perceptual system. Lady Macbeth had a first-order judgment of "dagger there" to accompany her experience of a dagger, despite the fact that there was no dagger to be perceived or experienced.

Even non-perceptual experience falls under this umbrella. Although there may be no *object* of a pain experience, contents along the line of "something hurts" are still cognitively represented. The very fact that we can comment on the hurt and direct our behavior appropriately brings out this fact. There is awareness here just as there is awareness in visual perception, even though the object of the awareness is not so clear-cut. A similar story goes for our experience of emotion, and other "internal" experiences.

Note that the principle is not that whenever we have a conscious experience we are aware *of the experience*. There is a sense in which that may be plausible, but that is a matter of a more indirect coherence between consciousness and second-order judgments, along the lines of the principle of detection. Rather, I am saying that when we have an experience, we are aware of the *contents* of the experience. When we experience a book, we are aware of the book; when we experience a pain, we are aware of something hurtful; when we experience a thought, even, we are aware of the contents of that thought. It is not a matter of an experience followed by a judgment, as might be the case for second-order judgments; these first-order judgments are concomitants of experiences, existing alongside them.

Although we have seen that experiences *can* be noticed and commented on, this is not the usual state of affairs. Most of the time we do not notice our experiences, but only the contents of our experiences. Only in a reflective or a philosophical mood do we sit back and take note of our *experience* of a red book; most of the time, we just think about the book. Second-order judgments are relatively infrequent, but first-order judgments are ubiquitous. The link between consciousness and first-order judgments is therefore more direct than the link with second-order judgments. (There are some hard cases, such as the fleeting contents of experiences that go by too quickly to be noticed. Even in these cases, though, it is likely that the contents are cognitively represented, even though they leave no trace.)

The arrow goes both ways. Where there is awareness, there is generally consciousness. When we are aware of something in our environment, with some reportable content directing our behavior, there is generally a corresponding conscious experience. When my cognitive system represents a dog barking, I have an experience of a dog barking. When I am aware of heat around me, I feel hot. And so on.

Not all the contents of awareness are contents of experience, though, at least as the notion of awareness stands. We have various kinds of awareness, especially those involving information in memory, that do not seem to have corresponding experiences. I am aware that Clinton is president, in the sense that I have access to this information, can verbally report it, and can use it in the deliberate direction of my behavior. There does not seem to be a corresponding conscious experience, though; or if there is, it is an extremely weak one. I am aware that there is a bicycle downstairs, without there being much of a corresponding bicycle-experience. This sort of awareness without experience is generally *propositional* awareness, but it need not be limited to that. It seems reasonable to say that I am aware *of my bicycle*, for instance; this is not propositional awareness, but awareness of an object along the lines that we found awareness in perception.

We could leave this matter as it stands, but it is more satisfying to try to put restrictions on the notion of awareness so that it is more truly parallel to consciousness. It does seem plausible that there is *some* kind of functional difference between the processes involved in one sort of case and in the other—the very fact that I can report on the difference between them bears witness to that. It is this functional difference that needs to be isolated.

Perhaps the most salient difference is that in cases of awareness with consciousness, there is a kind of *high-bandwidth* access that cases of awareness without consciousness lack. We can think of the bandwidth of informational access as proportional to (a) the *amount* of information accessible, and (b) the *strength* (or urgency) of the access relation, where (b) might be understood in terms of the extent to which the information conveyed tends to automatically occupy the resources of later processes. We can think of these by analogy to the amount of information and the strength of a radio signal. More usefully, note that the notion of awareness was explicated in terms of *access* (an "incoming" relation) and *control* (an "outgoing" relation). Both of these come in degrees. The bandwidth is a measure of the amount of access and the degree of control. (Speculations on the relation between bandwidth and consciousness have been put forward by Dennett 1991, although he uses these speculations for quite different purposes.)

When I am perceptually aware of a bicycle here in my room, I have direct access to all sorts of information about it, information that is pouring into my perceptual system and available to later processing *en masse*. In my awareness of the bicycle downstairs, by contrast, the awareness consists in limited access to information stored in memory: perhaps propositional information as limited as "bicycle down there", or perhaps something more, but in any case impoverished access compared to the perceptual case. My awareness in this case is extremely low-bandwidth, relative to the high-bandwidth access provided by perception.

In a similar way, the access to contents of pain experiences can be extremely high-bandwidth, as one would expect for a system whose function is to direct action immediately. Our access to some emotional contents may be low-bandwidth, corresponding to emotions that we are aware of but that are experienced only faintly; others are more all-consuming, and these are the ones that are salient in experience. An intense pain and even an intense emotion may not convey *much* information, but the information is conveyed strongly—it is sort of thing to which later processes are automatically sensitive.

All this is quite speculative, and would better be borne out with the aid of neurophysiological data, but it is at least a plausible hypothesis that consciousness corresponds to high-bandwidth access. We might even suppose that the intensity of a conscious experience varies directly with the bandwidth. It is not entirely implausible that even cases of low-bandwidth access, such as my awareness of my bicycle or of Clinton, have very weak conscious experiences associated them. These experiences are certainly not as obvious as perceptual experiences, but on introspection it does seem that there may be some faint phenomenal tinge. Given the lack of intensity of the experiences, it is not surprising that they should be barely noticeable.

The question of the functional correlates of consciousness is a complex one, probably best carried out by a combination of empirical methods and careful *a priori* reflection. For now I have downplayed the role of empirical methods, as it takes a good deal of *a priori* reflection to get us into the ballpark. It takes philosophical assumptions even to come up with appropriate empirical criteria for consciousness in others, and what such methods can tell us is therefore limited by those assumptions. To determine that a given mechanism is a correlate of consciousness, we need grounds for believing that there is indeed consciousness accompanying it, and that requires considerable reflection on just what sorts of processes are likely to be so accompanied in the first place. Empirical methods cannot therefore replace philosophical reflection, but they can be a very useful aid to the imagination as well as a constraint on speculation. And given that we might want a theory of consciousness to be foremost a theory of *human* consciousness, such methods are vital for the purposes of determining just what the associated mechanisms in humans *are*. Detailed speculation about the empirical details would be premature at this stage, however.

In any case it seems very likely that there is *some* functional constraint on awareness such that the appropriately constrained version of awareness correlates with consciousness. If my suggestion above is not perfect, then a related constraint might do better. For now, I will assume that something like what I have suggested will work; the following discussion should be adaptable to a better account.

6.3 The principle of structural coherence

So far we have a hypothesis: where there is consciousness, there is awareness, and where there is (the right kind of) awareness, there is consciousness. The correlation between these can be made more detailed than this. For instance, we have seen that intensity of consciousness may correlate with intensity (or bandwidth) of awareness.

More importantly, it seems that the detailed *structure* of a conscious experience is recoverable from the structure of awareness. My phenomenal field is not a homogeneous blob. My visual field, for instance, has a definite geometry to it. There is a large red patch, beside it a small yellow patch, with some white in between; there are patterns of stripes, squares and triangles and so on. In three dimensions, I have experiences of shapes such as cubes, experiences of something being behind another thing, and so on. My visual field is made up from a mass of such details, and vitally, all these details are cognitively represented. The size and shape of the patches is represented in my visual system—perhaps by a fairly direct topographic map in the visual cortex, but even if not, it is represented *somehow*, as demonstrated by the fact that the system provides information that guides behavior appropriately. The stripes are perceptually represented, as is the cubical shape. In short, all of the detailed structure of my visual field is cognitively represented, available in the deliberate control of behavior, and verbally reportable. The structure of my experience is entirely mirrored in the structure of my awareness.

The same goes for *implicit* structure in the phenomenal field, such as relations between colors. Even if I am only seeing one color at a given time, there are a host of colors I *could* have been seeing, colors to which this color bears a structural relation. One color is very similar to another color, and quite different from another. Two colors can seem complementary, or one color group can seem "warm" and another "cold". It turns out that all this structure in the phenomenology of color is mirrored in our perceptual system. This is not surprising given the fact that the information is available to direct behavior, but it is still interesting to see the details, many of which are presented in a thorough account by Hardin (1988). We might say that the difference-structure in our conscious experience is mirrored by a difference-structure in our awareness; to the manifold of color experiences and relations among them, there corresponds a manifold of representations, with corresponding relations.

A similar story holds for structure in other phenomenal domains. The phenomenological structure in a musical chord is mirrored by structure in what is represented. The difference between an intense pain and a mild tingle corresponds to a difference in representation. In general, any detailed structure that one might find in one's phenomenal field will necessarily be mirrored in the structure of awareness, as we can see from the fact that the structure is reportable. The reverse will also hold, if the notion of awareness is appropriately restricted as above: structure in the right kind of awareness will be mirrored by structure in consciousness.

We then have a further hypothesis of coherence between consciousness and awareness: structure in consciousness is mirrored by structure in awareness, and vice versa.[1] Alongside the general principle that where there is consciousness there is awareness (and vice versa) and the principle relating the intensities of these, this makes for a systematic and central relation between phenomenology and psychology. I will group these principles together under the name of the *principle of structural coherence.*

It might be thought that certain odd cases are counterexamples to these coherence principles[2] . For instance, blindsight—a pathology arising from damage to the visual cortex (see Weiskrantz 1986)—is sometimes taken to be a case where consciousness and the associated functional role come apart (Heil 1983). Subjects with blindsight can see nothing in certain areas of their visual field, or so they say. If one puts something – a red or green light, for instance—in their "blind" area, they claim to see nothing. However, when one *forces* them to make some choice about what is in that area—perhaps to choose between red and green—it turns out that they are right far more often than they are wrong. Somehow they are "seeing" what is in that area without really *seeing* it.

This is not a situation that one would expect intuitively, and some have put it forward as a counterexample to functionalist theories of consciousness. After all, in blindsight there is discrimination, categorization, and even verbal report, but it seems (it is claimed) that there is no *experience.* If so, then might this be a case of awareness without consciousness?

Such a conclusion would be ungrounded. For a start, it is not *obvious* that there is no experience there—perhaps there is a faint consciousness, or perhaps there is even a strong consciousness to which the reporting facilities do not bear the usual relation. Even if there is no consciousness, however, it is obvious that this is not a standard case of *awareness*, either. Clearly there is a vast difference between the functional roles played here and those played in the usual case—it is precisely because of this difference in functional roles that we notice something amiss in the first place! In particular, subjects with blindsight seem to lack the usual kind of access to the information at hand: their access is curiously indirect, it is *extremely* low-bandwidth, and it does not seem to be available for the control of behavior in the usual way.

---

[1]This is closely related to Jackendoff's Hypothesis of Computational Sufficiency: "Every phenomenological distinction is caused by/supported by/projected from a corresponding computational distinction" (Jackendoff 1987, p. 24).

[2]In the following discussion I am indebted to the discussion in Dennett 1991. See also Tye (forthcoming) for an interesting discussion of blindsight and related cases.

For this reason, it seems that there is little reason to suggest that blindsight patients are aware of the information, in the relevant sense of "aware". At best they have the kind of awareness that goes into propositional awareness, such as my awareness of Clinton being president; in fact, their awareness is even lower-bandwidth than this, as revealed by the weakness of their ability to report. Perhaps they qualify as aware in some very weak sense, but if so, it is not impossible that they are also experiencing the information in a correspondingly weak way. The situation is quite underdetermined, given our lack of access to the phenomenal facts of the matter, and the lack of correlates in our own everyday life, but there is no compelling reason to suppose that these cases count against the coherence between consciousness and awareness.

A similar response can be given to other unusual cases, such as hysterical blindness. In all these cases, the very fact that we use behavioral evidence in order to reach conclusions about consciousness suggests that such subjects are functionally quite different from the usual, and furthermore that they are functionally different in a way that makes consciousness seem less plausible. There is little chance that such cases could provide evidence *against* a link between functional organization and conscious experience, given that our interpretation of such cases *relies* on such a link. Close examination of each of these cases shows that the evidence for unusual states of consciousness in such cases is entirely reliant on evidence of unusual states of awareness. Such cases therefore do more to bolster the principle of coherence than to damage it.

The explanatory role of structural coherence

Some have been sufficiently impressed by the coherence between consciousness and cognition to suggest that this is all we need for a reductive explanation of consciousness. Van Gulick (1993), for instance, notes the fact that the structure of our color space is represented functionally, and suggests that this closes the "explanatory gap" by providing a functional explanation of color sensation.

Things are not that simple, as must be clear by now. Nothing in the structure of awareness can explain *why* there is conscious experience at all. The pattern of discrimination and reaction in color perception is logically compatible with the absence of any kind of sensation. It follows that the existence of sensations is a further fact, and that the explanatory gap remains wide open.

The structure of awareness can nevertheless be very useful in explanation. If we hold the principle of structural coherence as kind of background assumption, then a functional account of the structure of color awareness will explain the structure of color experience, relative to that assumption. But there are limits. First, it seems unclear that this method can explain the intrinsic nature of a color experience itself, although it can explain the relational structure *between* such experiences or between parts of a complex experience. Second and more importantly, an account of the structure of awareness cannot explain why the principle of structural coherence holds in the first place. By bringing in the principle as a background assumption we have already moved beyond the stage of *reductive* explanation; the principle simply assumes the existence of consciousness, and does nothing to explain it. Indeed, our knowledge of the principle is based entirely on our knowledge of consciousness.

Within these limits, the principle of structural coherence provides an enormously useful kind of explanatory relation between the physical and the phenomenal. If we want to explain some apparent structure in a phenomenal domain—say, the relations we find between our experience of musical chords—then we can investigate the functional organization of the corresponding psychological domain, taking advantage of insights from cognitive science and neuroscience to reductively explain the structure of awareness in that domain. In doing so we have explained the structure of the phenomenal domain, modulo the contribution of the principle of structural coherence. Because of our appeal to this principle we will not have explained consciousness itself by doing so, but we will still have explained much of what is special about a *particular* phenomenal domain.

It is even reasonable to suppose that this way we might get some insight into what it is like to be a bat! Functional organization can tell us much about the kind of information that a bat has access to—the kinds of discriminations it can make, the ways it categorizes things, the most salient properties in its perceptual field, and so on—and about the way in which it uses it. Eventually we should be able to build up a detailed picture about the structure of awareness in a bat's cognitive system. By the principle of structural coherence, we will then have a good idea about the structure of the bat's experiences. We will not know *everything* about what it is like to be a bat—we will not have a clear notion of the intrinsic nature of the experiences, for instance—but we will know a lot. An interesting paper by Akins (1993) about the mental lives of bats can be read as contributing to this project, although I think she intends it more ambitiously.

In a similar way, Cheney and Seyfarth's (1990) book *How Monkeys See the World* is put forward as an answer to a question like Nagel's, taking us inside the mind of another species. In effect the work uses the principle of structural coherence as a background assumption throughout, giving an account of certain functional processes and the structure of awareness that they entail, and inviting us to infer a corresponding structure of experience. Of course this does not answer Nagel's real worry, for the usual reasons, but it is nevertheless a striking achievement. We do not need to engage the ultimate mystery of consciousness every time we want to account for a specific phenomenal domain.

Coherence as a psychophysical law

It is natural to suppose that these principles of coherence have the status of *laws*.

If consciousness is always accompanied by awareness and vice versa in my own case, and in the case of all humans, one is led to suspect that something systematic is going on. There is certainly a lawlike correlation there, in my case and in the case of many others. We can therefore put forward the hypothesis that this coherence is a law of nature: in any system, consciousness will be accompanied by awareness and vice versa.

The same goes for the full-blown principle of structural coherence. The remarkable correlation between the structure of consciousness and the structure of awareness is much too specific to be coincidental. It is natural to infer an underlying law from these correlations: for any system, anywhere in spacetime, the structure of consciousness will mirror and be mirrored by the structure of awareness.

If these principles are indeed laws, then this makes a significant contribution to a theory of consciousness, and in particular to the understanding of the supervenience principles involved. So far, all we know is that consciousness arises from the physical somehow, but we do not know in virtue of what physical properties it so arises; that is, we do not know what properties enter into the physical side of the connection. Given the laws of coherence, we have a partial answer: consciousness arises in virtue of the functional organization associated with awareness. We can even arrive a a fairly specific understanding of parts of the supervenience relation by virtue of the principle of structural coherence: not only does consciousness arise from awareness, but the structure of consciousness is determined by the structure of awareness.

Of course this law will probably not be a *fundamental* psychophysical law. We can see this by noting that the high-level notion of awareness enters into it, with all its associated vagueness. The notion of structure is also not fully enough specified to be a candidate for participation in a fundamental law. But not all laws are fundamental laws. All sorts of laws hold in biological and high-level physical domains that are not fundamental laws—think of laws involving fitness or species, or even laws concerning temperature and pressure in thermodynamics. It may even be that the principles of coherence are not *strict* laws; there may be some exceptions around the edges, especially given the underdetermined nature of the concept of awareness.

Even if these laws are not fundamental, and even if they are not strict, they can nevertheless provide much insight into the psychophysical relation. These laws provide a strong *constraint* that any fundamental psychophysical laws must satisfy. A proposed theory of consciousness that does not have the principles of coherence as a consequence will be in trouble. Furthermore, reflection on just what these principles entail may allow us to reach more specific conclusions about the underlying fundamental laws. If a proposed fundamental psychophysical law is very simple, well-motivated, and has the principles of coherence as a consequence, then that may provide good reason to accept it.

Accepting the principles of coherence as laws would also help settle some key questions, such as the question of whether computers might be conscious. Awareness is a functional property, realizable in any physical medium as long as it has the right pattern of causal organization. A silicon system can be just as aware as a neural one, if it is organized appropriately. If the principles of coherence are correct, then consciousness arises from functional organization, and there is no reason why an appropriately-organized silicon-based computer should not be conscious.

Lawful principles of coherence would be a significant step toward a theory of consciousness, then. We must now consider: what grounds do we have for accepting these principles as laws?

The basic evidence for these laws comes from the correlations in our own cases: ultimately, for me, from those in my own case. The apparent correlations between awareness and consciousness in my own case are so detailed and remarkable that there must be something more than a mere chance regularity. There must be some kind of underlying law. The only question, then, is *what* law? This law must entail that in my own case, awareness is always be accompanied by consciousness and vice versa, and further that the structure will correspond. The principles of coherence I have put forward will do the job, but perhaps others might also. What other principles, in competition with this one, might suffice?

It is very plausible that some kind of awareness is *necessary* for consciousness. Certainly all the instances of consciousness that I know about are accompanied by awareness. There seems to be little reason to believe in any instances of consciousness *without* the accompanying functional processes. If there are any, we have no evidence for them, not even indirect evidence, and we could not in principle. It therefore is reasonable to suppose on the grounds of parsimony that wherever there is consciousness, there is awareness. If we are wrong about this—if for instance a static electron has the rich conscious life of a Proust—then at least we will never know about it.

The question of the *sufficiency* of awareness is more difficult. Given the necessity of awareness, our candidates for an underlying psychophysical law will all have the form: "awareness plus something gives rise to consciousness". If they do not have this form—if for instance they are put in much simpler terms—they must at least *entail* a principle of this form, in order to explain the regularities in my own case. The remaining question, then, is: What is the extra something, or is nothing extra required?

Call this hypothetical extra ingredient the *X-factor*. Either I am conscious in virtue of awareness alone, or I am conscious in virtue of awareness and the X-factor. The X-factor might consistently be any property, so along as it is possessed by me now, and preferably throughout my life. Perhaps the X-factor is a matter of nationality, and awareness gives rise to consciousness only in Australians. Perhaps it is a matter of location, and awareness gives rise to consciousness only within a hundred million miles of a star. Perhaps it is a matter of *identity*, and awareness gives rise to consciousness only in David Chalmers.

All of these laws would be compatible with my evidence, and would explain the correlation, so why do they all seem so unreasonable? It is because in each of these cases, the X-factor seems quite *arbitrary*. There is no reason to believe that consciousness should depend on these things; they seem to be irrelevant frills. It is not as if the X-factor plays a role in explaining any of the phenomena associated with consciousness. At least awareness might help explain our phenomenal judgments, which have a close tie to consciousness, so there is some reason to believe in a connection there. By contrast, each of these X-factors seems to be appearing out of nowhere. Why would the universe be such that awareness gives rise to consciousness in one person, and one person only? It would be a strange, arbitrary way for a world to be.

The same goes for more "plausible" X-factors that somebody might put forward seriously. An obvious candidate for such an X-factor might be cell-based biology, or even human neurophysiology. Certainly some people have supposed that consciousness is limited to beings with the right kind of biological make-up. In a similar way, some have suggested that consciousness arises from functional organization only when that organization is not implemented in a "homunculi-headed" manner, such as the Chinese nation. But X-factors like these are equally arbitrary. They only complicate the fundamental laws, without any added compensation. Why should the world be set up so that awareness gives rise to consciousness only in beings with a particular biology, or such that internal homunculi are ruled out? The hypotheses seem baroque, an irrelevant frill.

Why might someone believe in such an X-factor? I think such beliefs arise for a natural but misguided reason. There is a basic intuition that consciousness is something over and above functional organization. As we have seen, this is an intuition that I share—consciousness is a further fact, for which no functional organization is logically sufficient. There is also a natural tendency to believe that everything is physical, and that consciousness must be physically explainable one way or another. Faced with these two pressures, there is a natural reaction: we have to add something extra, and the extra something must be physical. Human biology is a natural candidate for that extra ingredient. In this way, it might be thought that we have bridged the gap from functional organization to human biology.

But this is quite misguided. The addition of biology into the picture has not helped the original problem at all. The gap is as large as ever: consciousness seems to be something over and above functional organization and biology. As argued earlier, *no* physical facts suffice to explain consciousness. The X-factor can do no work for us; we are looking in the wrong place for a solution to our problem. The problem was the assumption of materialism in the first place. Once we accept that materialism is false, it becomes clear that the search for a physical X-factor is irrelevant; instead, we have to look for a "Y-factor", something *additional* to the physical facts that will help explain consciousness. We find such a Y-factor in the postulation of irreducible psychophysical laws. Once we have imported these into our framework, the intuition that consciousness is a further fact is preserved, and the problem is removed

The desire for a physical X-factor is a holdover from the attempt to have one's materialist cake and eat one's consciousness too. Once we recognize that consciousness is a further non-physical fact and that there are independent psychophysical laws, the X-factor becomes entirely redundant. To ask for an independent psychophysical connection *and* an X-factor is to ask for two gifts when we only need one.

The X-factor, then, has no explanatory role to play in a theory of consciousness, and only complicates the story. Any such factor will only end up making the fundamental laws more complex than they need to be. Given the simplicity of the picture on which awareness gives rise to consciousness, a universe in which consciousness depends on a separate X-factor begins to look like an unreasonable place. One might as well have a clause in Newton's laws saying every action has an equal and opposite reaction unless the objects involved are made of gold. Principles of simplicity dictate that the best hypothesis is that no X-factor is required and that awareness gives rise to consciousness *simpliciter*.

If we allow that there is an X-factor, then anything goes. No X-factor is better motivated than any other. Given that consciousness is to depend on something arbitrary, then for all we know it depends on something like size, skin color, or location. In such a situation, all bets are off, and the quest for a theory of consciousness would end here. It seems much more reasonable to go for the simplest powerful principle that explains the evidence, and assert the principle of structural coherence as a law. I cannot prove that there is no X-factor any more than I can disprove solipsism, but by far the cleanest and most satisfying picture leaves it out. As always, in developing a theory of consciousness, all we have to go on are principles of simplicity and plausibility, and the omission of an X-factor is the simplest alternative by far.

It is worthwhile taking a moment to grasp the overall epistemological framework. What we have here is essentially an inference to the best explanation. We note remarkable regularities between consciousness and awareness in our own case, and postulate the simplest possible underlying law. This is the same sort of reasoning that goes on in formulating physical theories, and even in combating skeptical hypotheses about causation and about the external world. In all these cases, the underlying assumption is that the world is a simple and reasonable place. Failing such an assumption, anything goes. With such an assumption, things fall into place.

It also seems, incidentally, that this is as good a solution to the problem of other minds as we are going to get. We note regularities between experience and physical or functional states in our own case, postulate simple and homogeneous underlying laws to explain them, and use those laws to infer the existence of consciousness in others (recall the "epistemological myth" of Chapter 2). It seems to me that reasoning like this implicitly underlies our belief in other minds in the first place; all I have done here is bring it into the open.

The position we are left with, then, is a kind of functionalist dualism. Consciousness is not functionally definable, and it does not *logically* supervene on functional organization, but it nevertheless *nomically* arises from the proper functional organization. It turns out that systems made of other materials but organized so that they have the same structures of awareness that we do will have the same kind of conscious experience as we do. That is, cases of "absent qualia"—where a functional replica of a conscious being lacks conscious experience—are empirically impossible, although logically possible.

Some people will still be unsure about this. It is true that the argument from X-factors is somewhat tentative and relies strongly on simplicity assumptions. I will give more concrete arguments for the same conclusion in the next chapter, using thought-experiments to make the case that a functional replica of a conscious being will have precisely the same kind of conscious experience.

It is interesting to speculate on just what our principles of coherence imply for the existence of consciousness outside the human race, and in particular far down the biological "chain". The matter is unclear, as our notion of awareness is only clearly defined for cases approximating human complexity. For instance, we have used an appeal to verbal reports in the definition, so that we are aware of some contents if they are verbally reportable. But this appeal to verbal reports is only a heuristic. If someone lacks verbal facilities, we do not thereby doubt that they are conscious. Instead, the ability to produce verbal reports is used as a kind of rough measure of the degree to which information permeates as system and is available to later processes. The basic *concept* of awareness is not founded on reportability, but on more fundamental notions of access and control.

It seems entirely consonant with our notion of awareness that a dog is aware, and even that a mouse is aware. These may or may not be *self*-aware, but that is not required for awareness. They perceive their environments in a way that is not entirely different in kind from the way we do, even if the processes involved are simpler. It seems reasonable to say that a dog is aware of a fire hydrant in the basic sense of the term "aware". The dog's central systems certainly have access to information about the hydrant, and can use it to control behavior appropriately. By the principle of structural coherence, it seems likely that the dog *experiences* the hydrant, in a way not unlike our visual experience of the world. Presumably we all believe that dogs have visual experiences, just as we believe in other minds; again, all I am doing here is bringing the assumptions in our reasoning to the foreground, and placing them on a slightly firmer foundation.

The same is arguably true for mice and even for flies. Flies have some limited perceptual access to environmental information, and their perceptual contents presumably permeate their cognitive systems and are available to direct behavior. It seems reasonable to suppose that this qualifies as awareness, and that by the principle of structural coherence there is some kind of accompanying experience.

Around here the matter gets tricky. It is tempting to extend the coherence further down the information-processing scale; but sooner or later, the notion of "awareness" gives out on us and can do no explanatory work, due to its indeterminacy. I will therefore not speculate further on the matter for now, but I will return to it later when we have more tools at our disposal.

6.4 Explanatory coherence

With the principle of coherence between consciousness and awareness under our belts, it is natural to use this as a basic principle from which other coherence principles can be derived. In particular, it may be that the principle of structural coherence can help explain the principles concerning second-order judgments, those of reliability and detection.

I will not complete that project here, but I will give a flavor of how it might go. As we have seen, the principle of structural coherence is basically a principle of correlation between the contents of experience and of first-order judgments. To derive the principles connecting consciousness with second-order judgments, it is natural to look for a connection between first- and second-order judgments to do the work. If there were a principle that held that first-order judgments were always at least potentially accompanied by corresponding second-order judgments, and that second-order judgments generally are the result of some first-order judgment, then this in combination with the principle of structural coherence would yield the principles of reliability and detection for free. This explanation of the connection between first-order and second-order judgments is entirely a matter of psychology, of course. No dubious assumptions about phenomenology are needed. Judgments are in the domain of psychology, as is the connection between them.

A connection between first-order and second-order judgments is not too hard to understand. First let us deal with the reverse direction, from second-order to first-order. Given a system making a second-order judgment such as "I am having a red experience", it is clear that information about something's being red is represented internal to the system. It is precisely because this information is represented that the system comes to judge that it has an experience in the first place. It follows that the system is also aware of the *contents* of the experience, and in effect has made a first-order judgment along the lines of "red object there". A judgment about an experience without an accompanying representation of the contents of the experience would be an odd situation. Most naturally, then, any second-order judgment will go along with a corresponding awareness of contents, or a first-order judgment.

To deal with the other direction, consider a system making a first-order judgment along the lines of "red object there". Given a reflective system with a capacity to introspect, one would expect a judgment of the form "I am aware of the red object" to arise, at least potentially. Given this direct awareness and this *knowledge* of direct awareness, it is natural to wonder about the fact of the awareness. As we saw in section 5.4, in a naturally designed system such awareness would simply throw it into one state or the other; from the system's point of view, its contents being one way rather than the other would simply be a brute fact, or a "qualitative" fact. Given all this it would naturally judge that it was "experiencing" the redness directly; this is the best way for the system to make sense of its situation. In this way, we can see that it is natural that a first-order judgment like "red object there" will have the potential to be accompanied by a second-order judgment like "I am having a red experience", if the system reflects on its awareness.

There remains also the question of explaining our third-order judgments, such as "Consciousness is mysterious". This is obviously a more difficult matter, as it requires an explanation of some fairly deep rational processes. It is unlikely that any simple account to do those processes justice. Still, the beginnings of this story might be seen in the account above, and that given in 5.4. From the system's point of view, the fact that it is in one state or another is simply a brute fact, a fact without explanation. It is natural for it to note "qualitative" differences, and to wonder where those differences come from. In this way we can see why awareness might seem mysterious to the system.

(Of course, this does not explain consciousness away. At best it explains certain judgments and claims. It is solely a story about processing, and as such leaves the questions experience itself untouched. Even our talk of a "point of view" here is only a heuristic device to explain judgments, and does not tell us anything about points-of-view in the stronger sense of subjects of experience.)

More needs to be said to flesh out these stories appropriately, but that is a matter better suited to empirical investigation than to armchair speculation. Even with the tenuous just-so stories given here, we can see why a connection between first-order and second-order judgments is a natural thing to expect in a cognitive system. Second-order judgments will almost always be accompanied by corresponding first-order judgments, and every first-order judgment will have the potential to produce a second-order judgment.

If an explanation likes this succeeds, what we have produced is a reductive explanation of some of our judgments about consciousness, along the lines indicated earlier. Furthermore, if we take all this in conjunction with the principle of structural coherence, we have explained something else: namely why those judgments are generally *right*. If we reductively explain why second-order judgments are accompanied by first-order judgments, conjoining this too the principle of structural coherence explains why second-order judgments have corresponding experiences to be *about*. Given (a) that judgments "I am aware of X" are usually right, (b) that these go along with second-order judgments "I am experiencing X", and (c) the structural coherence principle that awareness of X goes along with experience of X, it follows that a second-order judgment "I am experiencing X" will usually be accompanied by an experience of X. Of course there is more to say here, but this gives a sketch of the framework. In this way, it seems that we come as close to a satisfying explanation of our phenomenal judgments as we can in a nonreductive theory of consciousness.

In this way we also satisfy an important desideratum on a theory of consciousness, the requirement of *explanatory coherence*. We have seen that there are two key explananda involved with this subject matter. First there is the question of consciousness, which requires a non-reductive explanation. Second there is the question of our phenomenal judgments, which will have a reductive explanation. It would be a strange, arbitrary situations if these two explanations turned out to be entirely independent of each other. It would mean that the world was put together by a series of odd coincidences, producing phenomenal judgments that just happened to be right. The principle of explanatory coherence requires that a (non-reductive) explanation of consciousness and a (reductive) explanation of phenomenal judgments *cohere* with each other, and provide an explanation of why the judgments are generally reliable. This is just what an account like that above can provide.

The key to the explanatory coherence of the account I have given is the principle of structural coherence, which provides the key nexus between consciousness and cognition, and in particular between consciousness and phenomenal judgments. This principle links the reductive theory of the judgments with the nonreductive explanation of consciousness, and leaves the mind in surprisingly unified shape.[3]

A further desideratum might be an explanation of the principle of structural coherence itself, telling us in more fundamental terms just *why* that principle is true. Perhaps we might eventually move toward a double aspect theory, where consciousness and awareness are seen to be two different aspects of the same thing, with this double aspect being a consequence of some more basic double aspect at the fundamental level. In this way the mind would truly be brought into unification. This is a tall order in the primitive state of our understanding, but I will make a first attempt at such a fundamental theory in further work summarized in the appendix.

[3]The principle of explanatory coherence also provides further arguments against the "X-factor" hypothesis from earlier. If consciousness arises from an X-factor, such as biology or spatial location, that played no explanatory role with respect to our phenomenal judgments, then an overall theory of mind would once again be quite fragmented. The biological factors central to the explanation of consciousness would be quite independent of the functional factors central to the explanation of phenomenal judgments.

6.5 The epistemology of conscious experience

There is still some unfinished business. Although we have given an account of the coherence between consciousness and cognition, and in particular demonstrated an intimate link between an explanation of consciousness and an explanation of our phenomenal judgments, it remains the case that consciousness is explanatorily irrelevant to the judgments. Although our explanation of the judgments coheres with the facts about consciousness, it remains entirely reductive. The question therefore remains: how can this be squared with the fact that we *know* about consciousness?

Another way to see the problem is to note that not only will my zombie twin say the same things and form the same judgments about consciousness as I do, but that my claims and judgments will be formed by the same *mechanisms* as his. If *justification* accrues to these judgments solely in virtue of the mechanisms by which they are produced, as is frequently supposed, then it seems that the zombie will be as justified in his phenomenal judgments as I am. But on the face of it the zombie's judgments are not justified at all—after all, his claims are utterly and systematically false, and his claim to be conscious is something of a delusion. If my own judgments are no more justified than his, then I cannot claim to have *knowledge* of conscious experience. But this is crazy: if we know anything, we know that we are conscious. So something must have gone wrong somewhere. An opponent might claim that the error lies precisely in the nonreductive premise that zombies are logically possible in the first place.

These arguments have a certain force to them, but I think they are ultimately misleading and do not succeed. I will come at the matter indirectly, first considering a pair of replies that do not succeed in order to illuminate a correct approach.

The first obvious reply to the above is to make an analogy with our knowledge of the external world. This knowledge has sometimes been questioned on the grounds that there could be a creature to whom the world seemed exactly the same as it does to me—indeed, who might be a physical replica of me—but whose external world is nothing like I supposed. Think of the famous "brain in a vat", for example. This might be physically identical to my brain, but it is being fed artificial input signals by a team of neuroscientists. Despite the possibility of a brain in a vat, we are nevertheless generally prepared to say that *our* beliefs in an external world are justified. Might my relationship to a zombie twin be analogous to that with the brain in the vat, so that my belief is justified despite the possibility of the zombie?

The analogy is not compelling, unfortunately. An objector might argue that the mechanisms by which beliefs about the external world are formed *differ* between my case and the case of the brain in the vat. The internal mechanisms are the same, but the *external* mechanisms are quite different. My beliefs are produced by a certain kind of causal link with the environment, whereas his are produced by artificial signals. It might be argued that what justifies my beliefs is this external mechanism.

Certainly, the environment plays a central role in the *explanation* of my beliefs. By contrast, in the case of the zombie, *all* the mechanisms are exactly the same. Consciousness does not play an explanatory role in the formation of our phenomenal judgments analogous to the role played by the environment in the formation of our beliefs about the external world. (There is another relevant difference between the cases of the zombie and the brain in the vat that I will come to shortly, one that is more favorable to the nonreductivist about consciousness.)

An argument much like this has been made by Shoemaker (1975a) in an attempt to show that knowledge of consciousness requires materialism, and in fact reductive functionalism. Shoemaker bases his argument on a *causal theory of knowledge*, which holds that knowledge of some fact or object must bear some causal relation to that fact or object. Certainly that is the way that knowledge usually works—witness our knowledge of insects or of planets, or of the fact that grass is green. Shoemaker argues plausibly that our judgments about consciousness are physical and in fact functional states. From this it follows that if we have knowledge of conscious experience, and if the causal theory of knowledge is true, then consciousness must be physical. Shoemaker notes that if zombies (or the functional equivalents of zombies) were logically possible, then conscious experience could make no causal difference to our judgments about it, implying that knowledge of consciousness would be impossible.

For the purposes of this argument I will accept that consciousness is causally as well as explanatorily relevant, although I think there may be more to say here. I think the argument fails for a separate reason: the causal theory of knowledge is quite inappropriate for our knowledge of conscious experience.

Before getting to just what is wrong, I will consider another reply that a nonreductivist might make to these epistemological difficulties. A popular way to make sense of knowledge recently has been in terms of *reliability* (Armstrong 1973; Goldman 1986). Beliefs about a subject matter are justified if they are formed by a *reliable* process, that is, if the process by which they are formed tends to produce true beliefs. Perceptual beliefs are justified, for instance, if they come about through the process of optical stimulation from the environment, which generally produces true beliefs. They are not justified if they are produced by hallucination, which is a very unreliable mechanism. In the previous section, we have seen that the mechanisms by which our phenomenal judgments are formed is quite reliable. Given the principle of structural coherence, it turns out that these judgments will generally be true. It might be thought that such a reliabilist account of knowledge could be the answer to our difficulties, and that our phenomenal judgments constitute knowledge in virtue of the principle of reliability.

There is a certain appeal to this response, but it has a fatal problem: it cannot account for the *certainty* of our knowledge of consciousness. We have seen that phenomenal judgments need not be incorrigible, but my general judgment that I am conscious now does not seem to be the sort of thing about which there is any room for doubt. This is one matter about which Descartes was right.

Knowledge justified only by reliability, by contrast, will always be open to skepticism. We cannot be certain, after all, that a given belief is being produced by a reliable method. Even when I form a belief about the book on my desk by standard perceptual means, it is possible for all I know that the belief has been formed by other methods—by hallucination, or by neuroscientists—and that the belief *might* therefore be false. Justification based on reliability alone may account for *knowledge*, but it cannot account for *certainty*.

If the justification for my belief in consciousness were based on the reliable connection alone, then my belief that I am conscious now could not be as certain as it is. To see this another way, note that *even if* I were were a zombie until two minutes ago and all my past phenomenal judgments were false, then if things for me right now were as they are, I would still be utterly certain that I am conscious. My current evidential situation renders me certain even though it is compatible with the absence of a reliable mechanism. My justification cannot therefore be a matter of reliability alone.

Causal theories of knowledge have precisely the same problem. In cases where justification accrues to our beliefs solely in virtue of some causal connection, it is always possible that we might be wrong. Where there is causation, there is contingency. Even if there is a tight causal connection between our judgments about the world and the world itself, it is *possible* for all we know that the world is not out there. We might be brains in a vat, for instance. This is precisely because we cannot be certain about the causal link itself. Even if we could be, this would show that the basic justification for our belief lay not in the causal link, but in the justification for our belief in the link, this could not itself be a matter of causation.

The existence of alternative skeptical hypotheses goes along with any belief whose justification is merely causal. Our current evidential situation underdetermines its possible antecedents. Of course we need not *accept* the skeptical hypotheses—we have good grounds for believing that by far the most reasonable way for our evidence to have been caused is by the external world—but it at least remains a possibility not absolutely excluded by the evidence. The same goes even more strongly for particular pieces of causally mediated knowledge, such as my knowledge of a book on my table. There may well be a causal link between the book and my judgment, but for all I know it is *possible* that there is no book there, but merely a pattern of light.

With consciousness, however, there is no corresponding skeptical hypothesis. We know that we are conscious more directly than we know about the external world, or about books on a table. Someone might raise the possibility of the zombie, but this is no analogy at all. We know for certain that we are not zombies—our evidential situation rules it out directly. This, incidentally, is the second disanalogy between zombies and brains in vats mentioned earlier. We are not certain that we are not brains in a vat, but we are certain that we are not zombies. Our evidential situation rules out one possibility but not the other.

Someone might deny this certainty that we are conscious. If they did, we could retreat to the reliabilist account above, which works at least as well as a causal account. But the denial should not be accepted. If an alternative skeptical hypothesis—for instance the hypothesis that we are zombies—were consistent with our evidence, then we should accept that hypothesis. It would make things much easier; consciousness only seems to cause difficulties. But consciousness is *part* of our evidence, and it is not something that we can deny.

It follows that a causal theory of knowledge is quite inappropriate for consciousness. If justification accrued from a causal link, there would always be the possibility that the causal link was different from what we suppose, and that our evidence had arisen from some quite different cause. No causal theory can account for our utter certainty that we are conscious. It follows that justification for our belief in consciousness cannot be due solely to some causal link.

The objection to non-reductive theories based on causal theories of knowledge has therefore been dealt with. If we are certain that we are conscious, our justification must lie in something other than a causal link between our judgments and experience. This progress is only negative, however. We have defeated one view of the justification of our phenomenal judgments, but we need something to put in its place.

The solution may lie in the fact that conscious experience itself is constitutive of our justification. Conscious experience is part of our *evidence* about the way the world is—indeed, it is fair to say that it may be the most basic evidence of all. Our belief in consciousness is not something that requires *further* justification from something extrinsic. Our immediate acquaintance with conscious experience provides the justification.

To see this, consider the question of why it seems plausible that my zombie twin is not justified in his judgments about consciousness, despite the similarity of his mechanisms. The answer seems to be simply that he is not in the same evidential situation as me. He lacks certain vital evidence that I possess, namely conscious experience itself.

It follows that an account of the justification of our phenomenal beliefs cannot be given entirely in terms of the mechanisms by which they are produced. Looking at mechanisms alone, the zombie's beliefs are just as justified as mine. Or perhaps better, looking at mechanisms alone, my phenomenal judgments are just as unjustified as the zombie's. (We can see this for reasons independents of zombies: the justification that mechanisms provide will always be uncertain, due to the mechanisms' contingency.)

But if the non-reductive view is correct, then I am not just a set of mechanisms, and the justification for my judgments accrues from elsewhere. The justification of my judgments about consciousness is parasitic on consciousness itself, which is not a mechanism but is nevertheless constitutive of our epistemic situation.

The question remains: how can our judgments be justified if consciousness is explanatorily irrelevant to those judgments? The answer, again, is that justification of those judgments does not lie in their causal ancestry. Their justification lies in the fact that they are had by a system that is immediately acquainted with conscious experience.

It might be objected that if consciousness is explanatorily irrelevant, then this acquaintance does not play any role in the formation of the judgment. But that does not matter. The judgment is justified *due* to our acquaintance with consciousness, even if it is not *caused* by our acquaintance with consciousness. This may sound counterintuitive, but remember that what is primary to our epistemic situation here is our acquaintance, and that the judgment is only a secondary kind of outward manifestation, consisting in a disposition to make claims and the like. The judgment *inherits* its justification from my overall epistemic situation.

The justification of a phenomenal judgment accrues not from the mechanisms that produce it, as perhaps is the case with judgments about the external world, but from the overall state of the *person* having the judgment, including their inner states of acquaintance. My body and even my brain might not be not justified in their phenomenal judgments; it is *I* who am justified, and my claims and judgments inherit their justification from my overall state.

The reason why causal theories and the like are appropriate for most kinds of knowledge lies in the fact that those kinds of knowledge are causally mediated. Our knowledge of the external world is never absolutely direct, but is mediated by a complex process of physical causation. But as we saw in Chapter 2, conscious experience is at the very center of our epistemic universe. It has no distance to travel in order for us to know about it. It is simply there, constitutive of our very being. Because of its immediate nature, our acquaintance with consciousness is not to be understood by analogy with other kinds of knowledge, then. It is quite different in kind, sufficiently that the explanatory irrelevance of consciousness to our claims and judgments about it affects our knowledge of it not a bit.

There remains the question of what "immediate acquaintance" *is*, exactly. This question may be no more answerable than the question of what consciousness is. It is part of the nature of consciousness that subjects have some very direct relation to it. Indeed, it can be plausibly argued that conscious experience is partly *constitutive* of the subject. If so, the reason consciousness need not travel any distance to reach us is that it is *part* of us.

I do not pretend that this is perfectly clear. It is part of the mystery of consciousness and will need further unraveling. But at the least we can see that that are compelling reasons why belief in consciousness does not obtain its justification from some causal relation, or in terms of the explanatory role of consciousness. On the nonreductive view, our belief is justified because consciousness is part of us. I do not know consciousness in virtue of its contribution to my mechanisms, but in virtue of its contribution to *me*.

The paradox of phenomenal judgment reconsidered

Earlier we posed the paradox of phenomenal judgment, arising from the explanatory irrelevance of consciousness to our judgments about it. It seems to me that the paradox is now defused.

The first worrying thing about the paradox was that it seemed to make consciousness and cognition quite independent, leading independent lives. If so, it would seem to be a miracle that our phenomenal judgments are generally *right*. This view has been defused in section 6.4 by our acceptance of the principle of structural coherence. This principle not only makes the remarkable coherence between consciousness and cognition explicit, but it allows for an explanation of our phenomenal judgments upon which it becomes quite clear why these judgments are generally correct. The explanation of our judgments and the explanation of consciousness itself bear an intimate relation.

The second worrying thing was the thought that explanatory irrelevance would be incompatible with our knowledge of consciousness. But we have seen above we should never have expected our knowledge of consciousness to have been understood in terms of explanatory or causal relevance to our judgments. If consciousness is at the center of our epistemic universe, it does not matter that it cannot travel any distance; we know about it directly. The mechanisms by which our judgments are formed are irrelevant.

Perhaps the only real worry that explanatory irrelevance poses is that it seems remarkable and fortunate that the psychophysical laws are such that consciousness coheres so well with the physical mechanisms, with those laws working precisely in a fashion such that our phenomenal judgments come out right. To be sure, the principle of structural coherence explains the judgments, but is it not fortunate that the principle is true in the first place? After all, psychophysical laws are an independent variable, so to speak, and might have been otherwise. This suggest that we have not gotten to the bottom of the story, and that a fuller understanding of consciousness might allow us some understanding of why this sort of coherence is a property that we should expect the world to have. A double-aspect theory of consciousness might ultimately fulfill this role for us by showing how this coherence could be a consequence of a very simple, principled dual aspect at the most fundamental level of nature. I will pursue this sort of speculation in future work.

In any case, this is only an unanswered question, not something that threatens to fatally damage a nonreductive theory. It seems that any fatal damage has been averted, and that a nonreductive view centered on the principle of structural coherence can yield a surprisingly unified view of the mind.

Chapter 7

Absent Qualia, Fading Qualia, Dancing Qualia

7.1 Does functional organization determine conscious experience?

It is natural to suppose that consciousness *arises* one way or the other from the physical, whether or not it is physical itself. The question then suggests itself: in virtue of what sort of physical properties does consciousness arise? Presumably these will be properties of the brain, but it is not obvious just which properties are the right ones. Some have suggested biochemical properties; some have suggested quantum properties; many have professed uncertainty.

A popular suggestion is that consciousness arises in virtue of the *functional organization* of the brain. On this view, the chemical and indeed the quantum substrate of the brain is irrelevant to the production of consciousness. What counts is the brain's abstract causal organization, an organization that might be realized in many different physical substrates.

Functional organization can perhaps be best understood as the *pattern of causal interaction* between various parts of a system, and perhaps between these parts and external inputs and outputs. A description of the brain's functional organization will abstract away from the specific physical nature of the parts involved, and from the way that the causal connections in question are implemented. All that counts is the existence of an appropriate set of parts that can take on various different states, where the states of these parts stand in appropriate causal relations to each other. Two systems have a common functional organization if there is a correspondence between their parts and between states of their parts such that causal relations are preserved. For the purposes of the following discussion, I also stipulate that to realize a common functional organization at a particular time, two systems must actually be in corresponding states at that time; a specification of the functional organization of a system will include not just a specification of parts, states and relations between these, but a specification of the state that the system is *in*. I give a more formal account of functional organization using the notion of a *combinatorial state automaton* in Chalmers (1993a), but the informal understanding will suffice for our purposes here.

A system can have functional organization at many different levels, depending on how finely we individuate its parts and on how finely we divide the states of those parts. At a coarse level, the brain is divided into two hemispheres that interact in certain simple ways: when one hemisphere gets excited, so does the other, and when one is calm, so is the other. At this level, we can therefore see the brain as realizing a certain very simple functional organization that might be shared by a pair of spinning wheels connected by an axle that gradually transmits spin, or by a multitude of similar systems.

It is generally more promising, however, to view the brain's organization at a finer level. If we are interested in cognition, we will usually focus on a level fine enough to determine the behavioral capacities associated with the brain, where behavior is individuated to some appropriate level of precision. Organization at too coarse a level (e.g., the two-part/two-state level above) will fall far short of determining our behavioral capacities; the mechanisms that determine our behavior will fall through the cracks of this coarse description. The spinning-wheel system certainly does not share our behavioral abilities. At a fine enough level, however—the neural level, say, though perhaps a higher level would suffice—functional organization will determine behavioral capacities. Even if our neurons were replaced with silicon chips, then as long as these chips had an appropriate space of states with the same pattern of causal interactions as we find in the neurons, the system would produce the same behavior.

I claim: conscious experience arises from functional organization. More specifically, I claim that the brain has some functional organization F such that any realization of F will be accompanied by conscious experiences type-identical to those that accompany the brain. Whether F is realized in silicon chips, the Chinese population, or beer cans and ping-pong balls does not matter. As long as the functional organization is right, the conscious experience will be right. Call this the *invariance thesis*, as it holds that conscious experience will be *invariant* over realizations of the appropriate functional organization.

Of course this does not holds for *every* organization F that the brain realizes. The two-part/two-state organization is not responsible for consciousness; it might be realized in two spinning wheels without much in the way of accompanying experience at all. The claim is merely that for some sufficiently fine-grained organization, the invariance thesis will apply. To find a suitable fine grain, we must go down to a level that can capture the mechanisms that determine our behavior and behavioral dispositions. For the purposes of illustration, I will generally suppose that this is the neural level, although a higher level might suffice, and it is not impossible that a lower level could be required.

The thesis in question has generally been associated with a reductive functionalist view about consciousness: that is, the view that all it *means* to be conscious or all it *is* to be conscious is to be in the appropriate functional state, playing the appropriate causal role. From such a view the invariance thesis would naturally follow, but the invariance thesis can be held independently. Just as one can believe that consciousness arises from a physical system but is not a physical state, one can believe that consciousness arises from functional organization but is not a functional state. The view that I advocate has this form. It is a perhaps unorthodox combination of functionalism and dualism.

I will not be especially concerned with the dualistic aspects of my view below, mostly being concerned to argue for the invariance thesis. My arguments might even be embraced by materialist functionalists. They do not establish the full reductive functionalist conclusion, but they nevertheless support that position against other reductive views (say, a view on which consciousness is seen as a biochemical property). Of course, I think that all reductive views ultimately fail, but the following discussion will mostly be independent of that issue.

I have already argued for a version of the invariance thesis in Chapter 6, although not in so many words. In arguing for the principle of structural coherence—the principle that consciousness arises from awareness, without there being any extra "X-factor"—I have implicitly argued that consciousness arises from functional organization. Awareness is defined in functional terms, and is certainly invariant over realizations of sufficiently fine-grained functional organization. The arguments there were indirect and somewhat tenuous, however. In this chapter, I will use thought-experiments to argue for the thesis in a much more direct way.

Absent qualia and inverted qualia

The invariance thesis is far from universally accepted. Many people, of both dualist and materialist persuasions, have argued against it. Many have held that consciousness is associated only with systems having the right biological makeup, and that a metallic robot or a silicon-based computer could never be conscious. Others have conceded that a robot or a computer might be conscious if it was organized appropriately, but have held that nevertheless it would have a different kind of experience from the kind that we have.

There have generally been two kinds of argument against the invariance thesis. The first kind comprises arguments from *absent qualia*. In these arguments, a particularly bizarre realization of our functional organization is exhibited—a popular example due to Block (1980a) is a case in which our organization is realized in the population of China (as in Chapter 3). Surely, it is argued, *that* could not give rise to conscious experience. Therefore consciousness cannot arise from functional organization.

The second kind comprises arguments from *inverted qualia*, or from the inverted spectrum. According to these arguments, if our functional organization were realized in a slightly different physical substrate, a system might still have experience, but it would have a different kind of experience. Where we have red experiences, it might have blue experiences, and so on. Often these arguments are made in complex ways involving brain surgery, so that we wake up one morning seeing blue instead of red even though our functional organization is unchanged.

I have used arguments from absent qualia and inverted qualia myself (in Chapter 3), but I have appealed only to the *logical possibility* of absent qualia and inverted qualia. These phenomena seem straightforwardly logically possible to me; there is no contradiction involved in their description. Whether they are *nomically* or *empirically* possible is a different matter, however. It is logically possible that a plate may fly upward when one lets go of it in a vacuum on a planetary surface, but it is nevertheless empirically impossible. The laws of nature forbid it. In a similar way, establishing the logical possibility of absent qualia and inverted qualia falls far short of establishing their empirical possibility.

The logical possibility of absent and inverted qualia is sufficient to establish the failure of *reductive* functionalism, on which conscious experience is itself taken to be a functional property. But this is insufficient to refute the invariance thesis, which is a thesis about the connection between consciousness and the functional in our world. To refute the invariance thesis, the *empirical* possibility of absent or inverted qualia has to be established.

Many of those arguing for the possibility of absent or inverted qualia have been arguing only for logical possibility. They are exempted from the arguments that follow. Many have been arguing for an empirical possibility, however, and it is those arguments that I will address. (When the word "possibility" is used alone in what follows, it is always empirical possibility that is meant.)

In what follows, I will discuss arguments that have been put forward in favor of the empirical possibility of absent and inverted qualia, and will then offer detailed arguments *against* those possibilities. These arguments will crucially involve thought-experiments. Against the possibility of absent qualia, I will offer a thought experiment concerning *fading qualia*. Against the possibility of inverted qualia, I will offer a thought-experiment concerning *dancing qualia*.

These arguments from thought-experiment are only plausibility arguments, as always, but I think they have considerable force. To maintain the empirical possibility of absent and inverted qualia in the face of these thought-experiments requires accepting some implausible theses about the nature of conscious experience, and in particular about the relationship between consciousness and cognition. In this way, the invariance thesis will be established as the most plausible hypothesis.

7.2 Absent Qualia

Positive arguments for the empirical possibility of absent qualia have not been as prevalent as arguments for inverted qualia. Arguments for absent qualia almost all have the same form. Block (1980a) has made perhaps the most detailed presentation of this sort of argument.

These arguments usually consist in the exhibition of a realization of our functional organization in some unusual medium. It is pointed out that the organization of our brain might be simulated by the people of China, or even mirrored in the economy of Bolivia. If we got every person in China to simulate a neuron (we would need to multiply the population by 10 or 100, but no matter), and equipped them with radio links to simulate synaptic connections, then the functional organization would be there. But surely, says the argument, this baroque system would not be *conscious*!

There is a certain intuitive force to this argument. Many people have a strong feeling that a system like this is simply the wrong sort of thing to have a conscious experience. Such a "group mind" would seem to be the stuff of a science-fiction tale, rather than the kind of thing that would really happen.

But there is *only* an intuitive force. This is certainly nothing like a knockdown argument. It may be intuitively implausible that such a system should give rise to consciousness; but it is equally intuitively implausible that a *brain* should give rise to consciousness! Who ever would have thought that this hunk of grey jelly would be the sort of thing that could produce honest-to-goodness experiences? And yet it does. Of course this does not *show* that the Chinese population could produce a mind, but it is a strong *defeater* for the intuitive argument that it would not.[1]

Of course, we would not *see* any conscious experience in such a system. But this is nothing new; we do not see conscious experience in anyone. It might seem that there is no "room" for conscious experience in such a system, but again the same appears to be true of the brain. Thirdly, we might *explain* the functioning of the system without invoking conscious experience, but again this is familiar from the standard case. Once we absorb the true force of the failure of logical supervenience, it begins to seem not much more surprising that the population of China could give rise to conscious experience than that a brain could do so.[2]

---

[1]Points like this have been made by Lycan 1987 and many others.

[2]Churchland and Churchland (1981) have objected to the "Chinese nation" arguments on the grounds that it would need to handle around $10^{30,000,000}$ inputs to the retina, and an even vaster number of internal states of the brain. The population simulation, requiring one person per input and one person per state, would therefore require vastly more people than a population could provide.

This objection overlooks the fact that both inputs and internal states are combinatorially structured. Instead of representing each input pattern (over 10^8 cells) with a single person, thus requiring 2^{10^8} people, we only need 10^8 people to represent the input as a structured pattern. The same goes

Some have objected to the invariance thesis on the grounds that the functional organization might arise by chance, in the Bolivian economy or even in a pail (Hinckfuss, quoted in Lycan 1987, p. 32). But this could only happen by the most outrageous coincidence.[3] The system would need to have over a billion parts each with a number of states of its own (say, ten each). Between these states there would have to be a vast, intricate system of just the right causal connections, so that given *this* state-pattern, then this state-pattern will result, given *that* state-pattern, then that state-pattern will result, and so on. To realize the functional organization in question, these conditionals cannot be mere regularities (where this state-pattern happens to be followed by that state-pattern on this occasion); they have to be reliable, counterfactual-supporting connections, such that this state-pattern will be followed by that state-pattern whenever it comes up. (I discuss this further in Chalmers 1993a and 1993b.)

It is not hard to see that about 10^{10^9} such conditionals will be required of a system in order that it realize the appropriate functional organization, if we suppose a division into a billion parts. The chance that these conditionals could be satisfied by an arbitrary system under a given division into parts and states will be on the order[4] of 1 in $(10^{10^9})^{10^{10^9}}$ (actually much less, as the requirement that each conditional be reliable further reduces the chance that it will be satisfied). Even given the freedom we have in dividing a system into parts, it is extraordinarily unlikely that such organization would be realized by an arbitrary system, or indeed by *any* system that was not shaped by the highly non-arbitrary mechanisms of natural selection.

Once we realize the highly constrained complex structure that functional organization imposes on a system, it begins to seem less unlikely that even the population of China could support conscious experience if organized appropriately. If we take our image of the Chinese population, speed it up by a factor of a million or so, and shrink it into an area the size of a head, we are left with something that looks a lot like a brain, except that we have little homunculi where a brain would have neurons. On the face of it, there is not much reason to suppose that neurons should do any

for internal states. We therefore need no more people than there are cells in the brain.

[3]Bogen 1981 and Lycan 1987 make the suggestion that such "accidental" situations would not have qualia, as qualia require teleology. This would have the bizarre consequence of making the presence or absence of qualia dependent on the history of a system. Better, I think, to concede that such a system would have qualia while pointing out just how unlikely it is that such a system could arise by chance.

[4]This figure comes from noting that there are 10^{10^9} possible choices for the consequent of each conditional, representing the global state into which the system transits. The chance that a given global state will transit into the correct following state is therefore 1 in 10^{10^9}. In fact it will be lower, as any given global state will be realizable by many different "maximal" states of the physical system, each of which is required to transit appropriately. There are 10^{10^9} such conditionals to be satisfied, so the figure above falls out.

better a job than homunculi in supporting experience.

Of course, as Block points out, we *know* that neurons can do the job, whereas we do not know about homunculi. The issue therefore remains open. The important point is that this sort of argument provides only very weak evidence that absent qualia are empirically impossible. A more compelling argument is required to settle the matter one way or the other. *Perhaps* it is correct to say, as Block does, that our intuitions throw the burden of proof onto one who holds that qualia are functionally invariant, although I doubt it. In any case, I will take up that burden in what follows.

A separate argument that is sometimes put forward for the empirical possibility of absent qualia stems from the phenomenon of blindsight. It is argued that blindsight patients are functionally similar to us in relevant ways—they can discriminate, report contents, and so on—but that they lack visual experience. Therefore the functional organization of visual processing does not determine the presence of absence of experience.

We have seen in Chapter 6 that there is a significant *difference* between processing in normal subjects and those with blindsight, however. These subjects lack the usual kind of direct access to visual information. If the information is accessible at all, the access is very low-bandwidth, and the information is certainly not available for the control of behavior in the usual way. Indeed, it is precisely because of the difference in the organization of their processing, as manifested in their behavior, that we notice anything unusual in the first place and are led to postulate the absence of experience. These cases therefore provide no evidence against the invariance thesis.

7.3 Fading Qualia

My positive argument against the possibility of absent qualia will be based on a thought-experiment involving the gradual replacement of parts of a brain, perhaps by silicon chips. Such thought-experiments have been a popular response to these arguments in the folk tradition of artificial intelligence and sometimes in print. The gradual-replacement scenario is canvased by Pylyshyn (1980), although without a systematic accompanying argument. Arguments not unlike the one I am about to give have been put forward by Savitt (1982) and Cuda (1985), although these develop the arguments in different ways and draw slightly different morals from the scenario.[5]

This "Fading Qualia" argument will not be my strongest and most central argument against absent qualia; that role is played by the "Dancing Qualia" argument, to be outlined in 7.5, which also provides an argument against the possibility of inverted qualia. However, the Fading Qualia argument is perhaps more natural, is strong in itself, and provides good motivation for the second more powerful argument.

[5]For some related fables, see also Harrison 1981a, 1981b.

The argument takes the form of a *reductio ad absurdum*. Assume that absent qualia are empirically possible. Then there could be a system with the same functional organization as a conscious system (say, me), but which lacks conscious experience entirely. Without loss of generality, assume this is because the system is made of silicon chips instead of neurons. I will show how the argument can be extended to other kinds of isomorphs later. Call this functional isomorph Robot. The causal patterns in Robot's cognitive system are just as they are in mine, but he has the consciousness of a zombie.

Given this situation, we can construct a series of cases intermediate between me and Robot such that there is only a very small change at each step and such that functional organization is preserved throughout. We can imagine, for instance, replacing a certain number of my neurons by silicon chips. In the first such case, only a single neuron is replaced. Its replacement is a silicon chip that performs precisely the same local function as the neuron.

Where the neuron is connected to other neurons, the chip is connected to the same neurons. Where the state of the neuron is sensitive to electrical inputs and chemical signals, the silicon chip is sensitive to the same. We might imagine that it comes equipped with tiny transducers that take in electrical signals and chemical ions, relaying a digital signal to the rest of the chip. Where the neuron produces electrical and chemical outputs, the chip does the same (we can imagine it equipped with tiny effectors that produce electrical and chemical outputs depending on the internal state of the chip). Importantly, the internal states of the chip are such that the input/output function of the chip is precisely the same as that of the neuron. It does not matter how the chip does this—perhaps it does it by a lookup-table that associates each input with the appropriate output, perhaps it does it by a computation that simulates the processes inside a neuron—as long as it gets the I/O dependencies right. It follows that the replacement makes no difference to the overall function of the system.

In the second case, we replace two neurons with silicon chips. It will be easiest to suppose that they are neighboring neurons. In this way, once both neurons are replaced we can cut out the intermediary and dispense with the awkward transducers and effectors that mediate the connection between the two chips. We can replace this by any kind of connection we like, as long as it is sensitive to the internal state of the first chip and affects the internal state of the second chip appropriately (there may be a connection in each direction, of course). Here we will ensure that the connection is a precise copy of the corresponding connection in Robot (perhaps this will be an electronic signal of some kind).

Later cases proceed in the obvious fashion. In each succeeding case a larger group of neighboring neurons has been replaced by silicon chips. Within this group of chips, the biochemical substrate has been dispensed with entirely. Biochemical mechanisms are present only in the rest of the system, and in the connection between chips at the border of the group and neighboring neurons. In the final case, every neuron in the system has been replaced by a chip, and there are no biochemical mechanisms playing an essential role in the cognitive system at all. (It is likely that biochemical units other than neurons, such as glial cells, may play a non-negligible role in the human brain. If so, we replace those too; presumably they will correspond to parts of the appropriate functional specification. For simplicity I will assume here that the relevant parts are all neurons, however.) What we have is essentially a copy of Robot.

We can imagine that in each case the silicon system is connected to a body and is sensitive to bodily inputs and produces motor movements in the appropriate way (with transducers and effectors at the surface of the body). Each system is therefore both functionally and behaviorally identical to me, and shares precisely my behavioral dispositions.

To fix imagery, imagine that as the first system I am having rich conscious experiences. Perhaps I am at a basketball game, surrounded by shouting fans, with all sorts of brightly-colored clothes in my environment, smelling the delicious aroma of junk food, perhaps suffering from a throbbing headache, and so on. Let us focus in particular on the bright red and yellow experiences I am having from watching the players' uniforms. The final system, Robot, is in the same situation, but by hypothesis is experiencing nothing at all.

Question: *What is it like to be the systems in between?* What, if anything, are they experiencing? As we move along the spectrum of cases, how does conscious experience vary? Presumably the very early cases have experiences much like mine, and the very late cases have little or no experience, or almost none, but what of the intermediate cases?

Given that the system at the other end of the spectrum (Robot) is not conscious, it seems that one of two things must happen along the way. Either (1) consciousness gradually fades over the series of cases, before eventually disappearing, or (2) somewhere along the way consciousness suddenly blinks out, although the preceding case had rich conscious experiences. Call the first possibility *Fading Qualia* and the second *Suddenly Disappearing Qualia.*

It is not difficult to rule out Suddenly Disappearing Qualia. This would require a massive discontinuity in the dependence of conscious experience on the physical. If the replacement of a single neuron could be responsible for the vanishing of an entire field of conscious experience, then consciousness would be an unreasonable phenomenon indeed. We could even imagine switching back-and-forth between a neuron and its silicon replacement, such that a field of experience blinked in and out of experience on command. While one cannot *disprove* that experience works this way, it is a most unlikely hypothesis.

This leaves Fading Qualia. To get some idea of how implausible this would be, consider a system halfway along the spectrum between me and Robot, after consciousness has degraded considerably but before it has gone altogether. Call this system Joe. What is it like to be Joe? Joe, of course, is functionally isomorphic to me. He *says* all the same things about his experiences as I do about mine. At the basketball game, he exclaims about the glaring bright red and yellow uniforms of the basketball players.

By hypothesis, though, Joe is not having bright red and yellow experiences at all. Instead, perhaps he is experiencing tepid pink and murky brown. Perhaps he is having the faintest of red and yellow experiences. Perhaps his experiences have darkened almost to black. There are various conceivable ways in which red experiences might gradually transmute to no experience at all, and probably even more ways that we cannot conceive. But presumably in each of these the experiences must stop being *bright* before they vanish (otherwise we are left with the problem of the Suddenly Disappearing Qualia). Similarly, there is presumably a point at which subtle distinctions in my experience are no longer present in an intermediate system's experience; if we are to suppose that all the distinctions in my experience are present right up until a moment when they simultaneously vanish, we are left with another version of Suddenly Disappearing Qualia.

For the sake of imagery, imagine that Joe sees a faded pink where I see bright red, with many distinctions between shades of my experience no longer present in shades of his experience. Where I am having loud noise experiences, perhaps Joe is experiencing only a distant rumble. Not everything is so bad for Joe: where I have a throbbing headache, he only has the mildest twinge.

The crucial feature here is that Joe is systematically *wrong* about everything that he is experiencing. He certainly *says* that he is having bright red and yellow experiences, but he is merely experiencing tepid pink. If you ask him, he will claim to be experiencing all sorts of subtly different shades of red, but in fact many of these are quite homogeneous in his experience. He may even complain about the noise, when he is only experiencing a distant rumble. Worse, on a functional construal of belief, Joe will even *believe* that he has all these complex experiences that he in fact lacks. In short, Joe is utterly out of touch with his conscious experience, and is incapable of getting in touch.

This seems to be vastly implausible. Here we have a being whose rational processes are functioning and who is in fact *conscious*, but who is utterly wrong about his own conscious experiences. Perhaps in the extreme case, when all is dark inside, it might be reasonable to suppose that a system could be so misguided in its claims and judgments—after all, in a sense there is nobody in there to be wrong. But in the intermediate case, this is much less plausible. We have here a sentient, rational being that is utterly incapable of forming correct judgments about its own experience. This implies a strong dissociation between consciousness and cognition. If this sort of thing could happen, then once again consciousness would be an extremely ill-behaved phenomenon.

To be sure, it is *logically* possible that Fading Qualia could be the case. There is no contradiction in the description of a system that is so wrong about its experiences[6] . But logical possibility and empirical possibility are different things. We have not the slightest reason to believe that this sort of case could happen in practice, and every reason to believe otherwise. It seems to be a central feature of consciousness that when a conscious being has experiences, it is at least capable of forming judgments about those experiences. Perhaps there are some cases where the rational processes in a system are strongly impaired, leading to a malfunction in the mechanisms of judgment, but this is not such a case. Joe's processes are *functioning* as well as mine—by hypothesis, he is functionally isomorphic. It is just that he happens to be completely misguided about his experience.

Of course there are various cases of fading qualia in everyday life. Think of what happens when one is dropping off to sleep; or think of moving slowly down the evolutionary chain from people to bacteria. In each case, as we move along a spectrum of cases, conscious experience gradually fades away. But in each of these cases, the fading is accompanied by an corresponding change in *functioning*. When I become drowsy, I do not believe that I am wide awake and having intense experiences (unless perhaps I start to dream, in which case I very likely *am* having intense experiences). The lack of richness in a dog's experience of color accompanies a corresponding lack of discriminatory power in a dog's visual mechanisms. These cases are quite unlike the case under consideration, in which experience fades while functioning stays just the same. Joe's *mechanisms* can still discriminate subtly different wavelengths of light, and he certainly *believes* that such discriminations are reflected in his experience, but we are to believe that his experience does not reflect these discriminations at all.

Searle (1992) discusses a thought-experiment like this one, and suggests the following possibility.

> ...as the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when the doctors test your vision, you hear

[6] Cuda 1985 claims that a description of systems with such mistaken beliefs is *senseless*. He offers no argument for this apart from the claim that if the description made sense, it would make sense to think that *we* are mistaken in such a way, which it clearly does not. But this seems fallacious. It makes *sense* to suppose that I could be mistaken in this way, in that the hypothesis is coherent; it is just that my epistemic situation shows me that the hypothesis is not *true* in my own case, because I have direct experience of bright red qualia and the like.

> them say, "We are holding up a red object in front of you; please tell us what you see." You want to cry out, "I can't see anything. I'm going totally blind." But you hear your voice saying in a way that is completely out of your control, "I see a red object in front of me." If we carry the thought-experiment out to the limit, we get a much more depressing result than last time. We imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same.

Searle effectively embraces the possibility of a system suffering from Fading Qualia, but suggests that it need not be mistaken in its beliefs about its experience. The system has true beliefs about its experience; it is just that these beliefs are impotent to affect its behavior.

It seems that this is one possibility that can be definitively ruled out. There is simply no room in the system for any new beliefs to be formed. Unless one is a dualist of a very strong variety, beliefs must be reflected in the functioning of a system—*perhaps* not in behavior, but at least in some process. But this system is identical to the original system (me) at a fine grain. There is simply no room for new beliefs like "I can't see anything", new desires like the desire to cry out, and other new cognitive states such as amazement. Nothing in the physical system can correspond to that amazement. There is no room for it in the neurons, which after all are identical to a subset of the neurons supporting the usual beliefs; and Searle is surely not suggesting that the silicon replacement is itself supporting the new beliefs! Failing a remarkable, magical interaction effect between neurons and silicon—and one that does not manifest itself anywhere in processing, as organization is preserved throughout—such new beliefs will not arise.

An organization-preserving change from neurons to silicon simply does not change enough to effect such a remarkable change in the content and structure of one's cognitive states. A twist in experience from red to blue is one thing, but a change in beliefs from "Nice basketball game" to "Oh no! I seem to be stuck in a bad horror movie!" is of a different order of magnitude. Such a major change in cognitive contents must surely be mirrored in a change in functional organization. Otherwise, cognition would float free of internal functioning like a disembodied Cartesian mind. If the contents of cognitive states supervened on physical states at all, they could do so only by the most arbitrary and capricious of rules (if this organization in neurons, then "pretty colors!"; if this organization in silicon, then "Alas!").

It follows that there is no reasonable way for Fading Qualia to happen. Fading Qualia require either a bizarre relationship between belief contents and physical states, or the possibility of beings that are massively mistaken about their own conscious experiences despite being fully rational. A much more reasonable hypothesis is that when neurons are replaced, qualia do not fade at all. A system like Joe, in practice, will have conscious experiences just as rich as mine. Our original assumption must therefore have been wrong. Even Robot, the system made entirely out of silicon, will have rich conscious experiences.

As I said before, this is only a plausibility argument. It is logically consistent to maintain that qualia would fade, or even suddenly disappear. But as always, almost *any* position about consciousness is logically consistent, and plausibility arguments are all we have to go on. In this case, the plausibility seems to be entirely on the side of the qualia not fading. Between the hypotheses (a) that a rational conscious being is generally right in its judgments about its experience, and (b) that the being in question is massively wrong and incapable of getting things right, then the first choice is clearly the more natural, other things being equal. This way, consciousness turns out to be a reasonable phenomenon. This conclusion preserves a principled coherence between consciousness and cognition. The alternative makes consciousness arbitrary and capricious, and makes for a strong dissociation between consciousness and cognition. If the alternative were true, the mind would be as disunified as the former Soviet Union.

The argument can be straightforwardly extended to other functional isomorphs. To deal with the case where the population of China implements my organization, we can construct a similar spectrum of cases between my silicon isomorph and the population. Perhaps we first gradually expand the silicon system until it is many square miles across. We also slow it down so that the chips are receiving inputs at a manageable rate. After doing this, we get people to step in one at a time for the chips, making sure that they set off outputs appropriately in response to inputs. Eventually, we will be left with a case where the entire population is organized as my neurons were, perhaps even controlling a body by radio links. (We may assume that the silicon system was moved out of the body at an early stage.)

At every stage, the system will be functionally isomorphic to me, and precisely the same arguments apply. Either conscious experience will be preserved, or it will fade, or it will suddenly disappear. The latter two possibilities are just as implausible as before. We can conclude that the population system will support conscious experiences, just as a brain does. We can even extend the argument to the case of Searle's (1980) "Chinese room". (I will discuss this in future work summarized in the appendix.)

We can do the same thing for any functionally isomorphic system, including ones that differ in shape, size, speed, physical makeup, and so on. In all cases, the conclusion is the same. If such a system is not conscious, then there exists an intermediate isomorphic system that (a) is conscious, (b) has faded experiences, and (c) is completely wrong about its experiences. Unless we are prepared to accept this massive dissociation between consciousness and cognition, the original system must have been conscious after all.

If Absent Qualia are possible, then Fading Qualia are possible. But I have argued above that Fading Qualia are almost certainly impossible. Therefore Absent Qualia are impossible.

I will now deal with various objections to the argument.

Objection 1: Neural replacement would be impossible in practice.

Those of a practical bent might not be impressed by this thought-experimental methodology. They might object that replacing neurons by silicon chips is the stuff of science fiction, not the stuff of reality. In particular, they might object that this sort of replacement would be impossible in practice, and that any conclusions that can be drawn therefore do not reflect the realities of the situation.

If "impossible in practice" means only that *we* could not perform the surgery in question, given current or future medical and engineering capacities, then this is not much of a problem. All that the thought-experiment requires is that the spectrum of cases stretching from me to my robot cousin be *nomically possible*—that is, that their existence would not violate any laws of nature. If they are, then we can legitimately draw conclusions about the laws of nature from the scenarios in question. What we are interested in is what kind of experience such systems would have *if* they existed. The question of whether we could actually construct such a system is irrelevant.

A more interesting sort of objection argues that the cases we have described are nomically impossible for some reason. Perhaps silicon simply lacks the capacity to perform the functions in the brain that a neuron performs, so that no silicon chip could be up to the task. Now, if it turned out that this inability ruled out the possibility of my silicon isomorph entirely, then the invariance thesis would not be threatened, as the opponent of the thesis would no longer have an unconscious functional isomorph to appeal to. However, perhaps the problem has something to do with the silicon/neuron mixture. The pure neural system and the pure silicon system might both be quite possible; it is only the intermediate systems that are ruled out.

It is difficult to see exactly what the problem could be. One source of problems might focus on the interface between the silicon chips and the rest of the system—the transducers and effectors of which I have spoken. Maybe there just would not be enough room for them in the tiny space a chip has available. After all, the effectors may have to store a reservoir of chemicals in order that they can be emitted when necessary. It is difficult to see this as a *principled* objection, however. First, it is not obvious that there is not enough room—a neuron manages to store the relevant chemicals! Second, we would not need a very large supply—the argument can make its point if the system is isomorphic to us for just a few seconds. Third, we can always run the thought-experiment by supposing an *expansion* of the system, perhaps to cells with much longer axons and dendrites, so that there was much more room to play around in. In any case, this objection seems to rest on very inessential features of cognitive system; it is hard to see that *this* could be the reason why the invariance thesis fails.

Another objection might hold that a neuron's behavior is not computable, so that no silicon replacement could do the job. There is no reason to believe that this is the case, and good reasons to believe otherwise. The low-level laws of physics as we understand them today seem to imply that the evolution of physical systems is computable—see Pour-El and Richards (1989) for a discussion of this. In any case, uncomputability of neural behavior would seem to be the sort of thing that, if it ruled out silicon replacement of a neuron, would also rule out the pure-silicon functional isomorph. It would therefore be irrelevant in arguing against the invariance thesis.

Objection 2: Some systems *are* massively mistaken about their experience.

This objection points to the existence of real cases in which subjects are mistaken about their experience. In subjects with blindness denial, for instance, subjects believe they are having visual experience despite the fact that they very likely have none at all.

In all such cases, however, there is a large impairment to cognitive functioning. We are no longer dealing with fully rational systems. In systems whose belief-formation mechanisms are impaired, anything goes. Such systems might believe that they are Napoleon, or that the moon is pink. My "faded" isomorph Joe, by contrast, is a fully rational system, whose cognitive mechanisms are functioning just as well as mine. In conversation, he seems perfectly sensible. We cannot point to any unusually poor inferential connections between his beliefs, or any systematic psychiatric disorder that is leading his thought processes to be biased toward faulty reasoning. Joe is an eminently thoughtful, reasonable person, who exhibits none of the confabulatory symptoms of those with blindness denial.

The cases are therefore disanalogous. The plausible claim is not that *no* system can be massively mistaken about its experiences, but that no rational system whose cognitive mechanisms are unimpaired can be so mistaken. Joe is certainly a rational system whose mechanisms are working as well as mine. So the argument goes through.

Objection 3: Sorites arguments are suspect.

A ubiquitous objection notes that this argument has the form of a Sorites or "slippery-slope" argument, and observes that such arguments are usually suspect. A typical example of such an argument observes that a million grains of sand form a heap; if we take away a grain of sand from a heap, we still have a heap; therefore even one grain of sand constitutes a heap, which is ridiculous.

This reaction is based only on a superficial reading of my argument, however. Sorites arguments generally gain their leverage by ignoring the fact that some apparent dichotomy is in fact a continuum: there are all sorts of vague cases between definite heaps and definite non-heaps, for instance. My argument, by contrast, explicitly accepts the possibility of a continuum, but argues that the intermediate cases are impossible for independent reasons.

The argument would be a Sorites if it had the form: I am conscious; if you replace one neuron in a conscious system by a silicon chip it will still be conscious; therefore an all-silicon system will be conscious. But this is not its form. It is true that the argument against Suddenly Disappearing Qualia relies on the impossibility of a sudden transition, but importantly it argues against sudden *large* transitions, from rich conscious experiences to none at all. This is implausible for reasons quite independent of Sorites considerations. (This is not to say that there are no Sorites arguments to be found in the ballpark; see Tienson 1987 for an example).

Objection 4: Similar arguments could establish behavioral invariance.

A fourth objection argues that the argument proves too much. If it establishes the functional-invariance thesis, a similar argument would establish a *behavioral*-invariance thesis. To do this, we would construct a continuum of cases from me to any behaviorally equivalent system. It would follow by similar reasoning that such a system must be conscious. But it is plausible that some systems, such as Block's (1981) giant lookup-table that stores outputs for every pattern of inputs, are not conscious. Therefore there must be a flaw in the argument.

This objection fails in two ways. First, my argument relied partly on the fact that a functionally isomorphic system will have the same cognitive structure as me, and in particular the same beliefs. This is what led us to the conclusion that the faded system Joe must be massively wrong in its beliefs. The corresponding point does not hold for behaviorally equivalent systems. A perfect actor need not have the same beliefs as me. Nor will the lookup-table; nor will any intermediate systems. These will work by quite different mechanisms.

Second, it is not at all obvious how one could get from me to an arbitrary behavioral isomorph by taking small steps and preserving behavioral equivalence throughout. How would one do this for the lookup table, for instance? Perhaps there are ways of doing it by taking large steps at once, but this will not be enough for the argument: if there are large steps between neighboring systems, then Suddenly Disappearing Qualia are no longer so implausible. With functional isomorphs, there was a natural way to take very small steps, but there is no such natural method for behavioral isomorphs. It therefore seems unlikely that such an argument could get off the ground. (Even if such a sequence exists, it will still run afoul of the point in the paragraph above.)

I think there is only one tenable way for the opponent of the invariance thesis to respond to this argument, and that is to accept the possibility of Fading Qualia, and the consequent possibility that a rational conscious system might be massively mistaken about its experience. This position is unattractive in its implication of a dissociation between consciousness and cognition, and the alternative seems much more plausible, but unlike the other objections it is not *obviously* wrong. The Dancing Qualia argument in section 7.5 will provide even more evidence against the possibility of absent qualia, however, so opponents of the invariance thesis cannot rest easily.

It should be noted briefly that a similar sort of argument could establish that systems with *similar* functional organization to conscious systems will have conscious experience. The invariance thesis taken alone is compatible with the solipsistic thesis that my organization and only my organization gives rise to experience. But one can imagine a gradual change to my organization, just as we imagined a gradual change to my physical makeup, under which my beliefs about my experience would be mostly preserved throughout, I would remain a rational system, and so on. For similar reasons to the above, it seems very likely that conscious experience would be preserved in such a transition.

7.4 Inverted Qualia

The arguments above have established that my functional isomorphs will have conscious experience, but they have not established that isomorphs will have the *same* sort of conscious experience. If what has gone before is correct, then functional organization determines the existence or absence of conscious experience, but it need not determine the nature of that experience. To establish that functional organization determines the nature of experience, we will have to establish that functional isomorphs with *inverted* qualia are impossible.

The idea of inverted qualia is familiar to most of us. I imagine that few have not wondered whether what looks red to one person looks blue to another, and vice versa. It is one of those philosophical puzzlers where at first one is not sure whether the idea even makes sense, and that even on reflection can be baffling.

The possibility of inverted qualia was apparently first put forward by Locke, in his *Essay Concerning Human Understanding* (Book Two, Chapter 32, Section 15):

> *Though one Man's Idea of Blue should be different from another's.* Neither would it carry any imputation of falsehood to our simple ideas, if by the different structure of our organs it were so ordered that *the same object should produce in several men's minds different ideas* at the same time; v.g. if the idea that a violet produced in one man's mind by his eyes were the same that a marigold produced in another man's, and *vice versa.* For, since this could never be known, because one man's mind could not pass into another man's body, to perceive what appearances were produced by those organs, neither the idea hereby, nor the names, would be at all

> confounded, or any falsehood be in either. For all things that had the texture of a violet producing constantly the idea which he called blue, and those which had the texture of a marigold producing constantly the idea which he called yellow, whatever those appearances were in his mind, he would be able as regularly to distinguish things for his use by those appearances, and understand and signify those distinctions marked by the names 'blue' and 'yellow', as if the appearances or idea in his mind received from those two flowers were exactly the same with the ideas in other men's minds.

Of course, Locke is not especially concerned with inverted qualia between functional isomorphs, but merely with inverted qualia between systems with similar behavior. It certainly seems that a conceptual possibility is being expressed (as I have argued in Chapter 3). The question for us is whether an *empirical* possibility is being expressed.

Even those who consider themselves materialists have often supposed that functional isomorphs might have different conscious experiences. (By "different experiences", I mean experiences of a different *type*, rather than numerically distinct experiences, of course.) For instance, it is often thought empirically possible that a functional isomorph of me with different physical make-up might have blue experiences where I have red experiences, or something similar. This is the hypothesis of inverted qualia. If it is true, then while the presence of conscious experience might depend only on functional organization, the nature of experiences would depend on physiological makeup, or some other non-functional factor.[7]

We have seen earlier that this position cannot be held consistently with materialism. If it is empirically possible that my functional isomorph would have inverted qualia, then it is logically possible. It is therefore equally logically possible that my *physical* isomorph would have inverted qualia, as there is no more of a *conceptual* connection from neurons to a specific kind of qualia than from silicon. It follows that the nature of qualia provides a further fact over and above the physical facts, and that materialism must be false. I will leave this point aside in what follows, however. The discussion will be independent of the truth of materialism or dualism.

The possibility of inverted qualia, or of the "Inverted Spectrum" as it is sometimes known, is sometimes objected to on the verificationist grounds that we could never know that anything different was going on, so that there could be no objective difference there (e.g., Schlick 1932). Obviously I do not accept these arguments, for reasons given earlier (to be a certain conscious experience is not to play a certain

[7]This position is mostly closely associated with Shoemaker (1982), but has also been advocated by Horgan (1984a), Putnam (1981), possibly Lewis (1978), and various others.

functional role). The inverted-qualia hypothesis expresses a logical possibility; the question is whether it is an empirical possibility.

As I discussed in Chapter 3, the hypothesis is also sometimes objected to on the grounds that our color space is asymmetrical, so that no inversion could quite map things appropriately (e.g., Hardin 1987, Harrison 1967; 1973). Some of the responses I made in Chapter 3 are still appropriate here, even though the question is now one of empirical possibility. In particular, we can still appeal to the possibility of a creature with a symmetrical color space, and ask whether it could have an inverted functional isomorph. From now on, I will ignore this worry. For the sake of argument I will grant that we have a symmetrical color space, and argue that inverted qualia are impossible in any case.

Discussion of inverted qualia can be confusing. When I say "blue experience", do I mean (1) what a *subject* calls a "blue" experience, (2) an experience caused by a blue object, or (3) what *I* call a blue experience? I choose the latter usage. Throughout my discussion, by "blue experience" I will mean the kind of experience that *I* call "blue", that I have when I see blue things like the sky and the sea, and so on. On this usage, then, it is conceivable that others (or even a future version of me) might have blue experiences caused by yellow objects, or objects they call "red", and so on.

Various arguments have been put forward for the possibility of inverted qualia. I will here be concerned to establish that they do not make their case, before moving on to positive arguments *against* the possibility.

1. Intrasubjective spectrum inversion

The first argument stems from a *good* argument for inverted qualia, but one that does not affect the invariance thesis. This is an argument for the possibility of qualia that are inverted with respect to *behavior*. The argument is due to Gert (1965), Lycan (1973), and possibly others. An antecedent of the argument is present in a discussion by Wittgenstein (1968).

The argument establishes the possibility that qualia could be inverted between subjects by first noting that qualia could be inverted *within* a subject. Although we might never have direct evidence for the first case, we could certainly have evidence for the second case. Imagine that I wake up tomorrow and the sky suddenly seems red, I seem to be bleeding blue blood, grass seems bright yellow, and so on. This I will take as good evidence that my qualia have been inverted. Furthermore, this inversion will be reflected in my verbal reports—"The world looks extremely bizarre today"—and others will have good evidence that an inversion has taken place in me.

We can even imagine a mechanism by which this might take place. We need only imagine that during the night a demon fiddled around with my visual system, rewiring it so that blue-wavelength stimulation will set off internal states previously associated with red wavelengths, and so on. In this way, when I wake up the next morning and look at the sky, it will send me into an internal state previously associated with red wavelengths, leading me to utter "That looks red!", and presumably giving rise to a red experience (if qualia are dependent on central states, as seems likely).

Given the possibility of this sort of rewiring, the possibility is raised that there could be somebody whose brain had been rewired in this way since birth. When they first saw the sky, their different wiring caused them to have what I would call a red experience; of course they learned to call the experience "blue". Although they apparently use color vocabulary in the same way as I do—in fact, they may be behaviorally identical to me— they nevertheless have systematically different experiences. Perhaps this is unlikely, but it is at least an empirical possibility. It follows that the possibility of inverted qualia with constant behavior is a reasonable one.

This does not, however, establish the possibility of inverted qualia with fixed functional organization. To see this, we need only note that the demon's rewiring *changed* my functional organization! Before, there was some internal state that was triggered by the sky and led to "blue" utterances; now, there is no such state. In "crossing the wires", my organization has been changed in a significant way. Indeed, the very fact that there is a noticeable change in my behavior in the resultant state shows that functional organization cannot have been preserved. If it had been preserved, no behavioral change would have been evident.

Putnam (1981) and Shoemaker (1982) have used precisely this example to argue *against* functionalist accounts of qualia. There is a sense in which this may be reasonable. The thought-experiment in question may count against a *coarse-grained* functional-invariance thesis, on which it is held that the same sort of experience will always arise from states that are triggered by blue things and lead to "blue" reports. The possibility of the person who had been in the rewired state since birth shows that this invariance thesis is false. But it does not count against a *fine-grained* functional-invariance thesis of the kind I have been advocating. The rewired system has a different functional organization, due to its different pattern of interconnections between states.[8] (Levine 1988 makes a similar point.)

[8]Actually, even the argument against the coarse-grained invariance thesis could be resisted. It would not be completely implausible to hold that a rewired-from-birth subject would have just the same experiences as me upon looking at the sky, in virtue of the similarity in causal role. The question then is: why does overnight rewiring produce different experiences where rewiring from birth does not? There could be two answers to this: (1) overnight rewiring does *not* produce different experiences, but just plays havoc with one's memory of what old experiences looked like, so that one *thinks* the world looks different even while it looks the same; (2) overnight rewiring produces different experiences, but this is due to interference from memory circuits, the difference in functional role ("that looks weird!"), and so on; the rewired-from-birth subject does not share these differences, and so might have the same experiences as the original subject. I will not adjudicate the issue, but all in all it seems more plausible to me that the rewired-from-birth case would have different experiences.

Putnam and Shoemaker both use this thought-experiment to draw the conclusion that qualia are dependent on physiological makeup, rather than depending just on functional state. But this conclusion cannot be drawn from the scenario in question. All that we can conclude is that we need to go to a finer level of functional organization.[9] If we assume that the relevant functional organization is at a grain fine enough to determine the mechanisms that cause our behavior, then no thought-experiment like this one can count against the functional-invariance hypothesis.

In a variant of this scenario (e.g., Block 1990b; Putnam 1981; Shoemaker 1982), the "rewired" subject undergoes a process of adaptation, learning to associate the new color of the sky with the word "blue", and even undergoes amnesia, forgetting that things ever looked different. If this happens to me, we are to suppose, my new system will be functionally isomorphic to my old system but my experiences will be different. The experience I have on looking at the sky now will be the kind of experience I had on looking at red roses before.

It is not obvious, though, that the system will be functionally isomorphic to the original configuration even after adaptation and amnesia. The rewiring will still be there, for instance, connecting an input state to a different internal state B where it was once connected to state A. Now, it might be argued that due to the adaptation process, the functional role that that state B plays will have changed precisely to the functional role that state A plays—it will lead to "blue" reports, associations with Picasso's early period, and so on—so that the new system could still be isomorphic to the old system with state B in the new system corresponding to state A in the old. But even so, it is implausible that mere adaptation and amnesia could cause state B to play *precisely* the functional role that state A once played, although it could be similar at a coarse level. This would require a massive change in brain organization. And if there were such a massive change, it is not so implausible that in the process, the experience associated with state B would gradually revert to that once associated with state A.[10] This case therefore cannot provide compelling evidence against the invariance thesis.[11]

[9] It might be objected that fine-grained functional organization comes to just the same thing as physiology, but this is not so. A fine-grained functional organization can in principle be realized in silicon, or even in the population of China; not so for a physiological description.

[10] Cole 1990 claims to have empirical evidence (!) that experiences would revert in this way.

[11] There are various ways the scenario could be elaborated. Perhaps we could run a version where states A and B are not just local states but states of a whole "central system" on which qualia might plausibly be thought to depend. In this case, though, (a) it is quite implausible that the organization of the whole central system would stay unchanged over the adaptation/amnesia process, and (b) if it did stay constant, with the only changes affecting the "earlier" and "later" stages of processing, then we certainly no longer have an argument against the invariance thesis! The *overall* functional organization of the system will have changed significantly (we can't simply make state B correspond to the earlier state A as we did before, as this correspondence will not go through at nearly fine

2. Inverted Earth

A related argument has been put forward by Block (1990b), using an example due to Harman (1982). Block describes an imaginary planet, Inverted Earth, as follows:

> Inverted Earth differs from Earth in two respects. Firstly, everything has the complementary color of the color on Earth. The sky is yellow, grass is red, fire hydrants and green, and so on. I mean everything *really* has these oddball colors. If you visited Inverted Earth along with a team of scientists from your university, you would all agree that on this planet, the sky is yellow, grass is red, etc. Secondly, the vocabulary of the residents of Inverted Earth is also inverted. If you ask what color the yellow sky is, they (truthfully) say "Blue!". If you ask what color the (red) grass is, they say "Green!". (p.62)

Block goes on to describe a situation in which scientists knock someone (say me) out and put "color-inverting" lenses in my eyes. These lenses have the effect that light from red things will stimulate my retina in the way that green things usually do, and so on. Next they take me to Inverted Earth. Here, I wake up and find that everything looks normal. The sky still looks blue to me, grass still looks green to me, and so on, despite the fact that the sky is *really* yellow and the grass is *really* red. In conversation with the inhabitants of the planet, nobody will notice anything amiss.

Block uses this scenario to argue convincingly against the claim that qualia can be analyzed as intentional properties, where for example a blue experience might be seen as a perceptual state that is about blue things. After some time on Inverted Earth, the experience you have on looking at the sky will be as blue as ever, but it will typically be caused by *yellow* things, and indeed, unbeknownst to you, your public-language term "blue" will begin to refer to the property possessed by *yellow* things in the environment. We therefore have a change in the intentional content of your states (these are now caused by and refer to yellowness in the environment), but the experience is the same as ever. Qualia are therefore not intentional properties, and are not even lawfully correlated with intentional properties.

Block also uses the scenario to argue that qualia are not (correlated with) functional properties, however. He notes that when I am having a blue experience on Inerted Earth, my internal state is caused by yellow objects. When my twin, who I left back at home and who has no inverting lenses, has a blue experience, his internal

enough a level—mechanisms within B will not map straightforwardly to mechanisms within A). True, the functional organization of the central system will be the same, but only in a way that supports the invariance thesis. Precisely the same experiences are associated with states of the central system as before—it is just that these states and experiences are now caused by different environmental inputs.

state is caused by blue objects. Block notes that when we are having the same experience, we are functionally inverted. *My* internal state is caused by yellow objects, causes me to manipulate yellow things in a certain way, and so on, whereas his is caused by blue objects, controls his response to blue things, and so on.

(Actually, *difference* in functional state with sameness in experience cannot refute the invariance thesis. However, a systematic difference like this one could easily provide evidence against it, and one might even adapt it to a case with relevantly similar functional state but different experience. I will therefore proceed on the assumption that systematic difference in functional state with sameness in experience of the kind suggested here needs to be refuted.)

This does not show that my twin and I are functionally inverted, however; at least we are not inverted in a way that threatens the invariance thesis. All it shows is that our states play a different *wide* functional role. The states supporting blue experiences in each of us interact with objects in the *environment* differently, due to the difference in our lenses. But our *internal* organization is just the same. When he sees a blue object and I see a yellow object, we go into just the same internal states. The invariance thesis holds that experience is determined by *internal* functional organization, so this example cannot cause trouble for it.

Block responds to an objection like this one by noting that we can move the "lenses" inward in the system, rewiring things at the optic nerve or in the visual cortex, for instance. In this case there would be an inversion in functional organization but still a sameness in experience between me (on Inverted Earth) and my twin on Earth. There are two responses to this. First, once again, a *difference* in organization with sameness in experience cannot alone refute the invariance thesis. Second and perhaps more important, the functional organization of the *central* systems of myself and my twin will still be the same, and it is plausibly this aspect of organization on which experience depends. What goes on at the periphery affects experience insofar as it affects the central system. Now, this claim might be dubious (I personally find it plausible), but note that insofar as it is dubious, the claim that my twin and I have the same experience is equally dubious! For insofar as experience is dependent on the periphery, then the peripheral difference in organization between me and my twin will cause us to have different experiences. Either way, the invariance thesis is undamaged.

A similar argument will work no matter where an inverting element is placed. If it is placed in a sufficiently central area, it will affect experience at the same time as it affects functional organization, so there is no damage to the thesis (when I look at yellow things and my twin looks at blue things, we will have *different* experiences). If it is placed in a peripheral area, it will not affect experience but it will equally not affect the relevant functional organization (when I look at yellow things and my twin looks at blue things, we will have the same experience and the same central functional organization). Block notes in response to a related point that the "central area" need not have any sharp boundary, but this makes no difference to the argument. We can draw the boundary widely: all that matters is that insofar as experience differs between me and my twin, the state of our central system will differ.[12]

In general, the considerations driving Block's intuitions that experiences are the same or inverted in a given case are dependent entirely on the intuition that it is central systems that make a difference to experience. We are asked to believe that experiences are the same in certain cases precisely because central processing is unaffected; we are supposed to believe that experiences differ in cases where central processing differs. There is no way that such arguments relying solely on the dependence of experience on central processes could refute the invariance thesis.

3. Retinal tuning

A similar argument is given by Seager (1991, pp. 39–41). Seager supposes that the cells in our retina are "tuned up" the electromagnetic spectrum so that they are sensitive to higher frequencies of light, but so that they play the same functional role in the system. Seager argues that such a person's capacities would be very different from our own—for instance, they would no longer be able to distinguish red-hot iron from cold iron, as their retinal cones would no longer respond to the low-energy photons from hot metal. They would fail many standard color-discrimination tests. But they would nevertheless have the same color experiences as we do, although those experiences would be associated with different objects.

Seager notes that they have the same experiences despite not being functionally isomorphic, so that any thesis of correlation between experience and functional state will be in trouble. The range of responses to this is as before. First, difference in functional state with sameness in experience cannot disprove the invariance thesis. Secondly, the functional difference between these systems is in their *wide* functional role, involving their interaction with the environment. Their *internal* functional organization (everything from retinal outputs in) is just as it was before. In particular, when the tuned-up system is having a red experience, it is in precisely the same internal state as the previous system would have been when having such an experience; it is just that this state is triggered by different stimuli. We can move the rewiring inward, but the moral of the Block discussion will again apply. Any reason

[12]Note that this talk of "central systems" is in fact a *concession* to our opponents rather than a defensive maneuver. We could stand fast and simply note that any internal change, even a peripheral one, is a change to functional organization, so that the invariance thesis is unthreatened; and we could suggest that maybe the internal change in fact *affects* the nature of the experience. The talk of "central systems" is simply an attempt to take seriously our opponents' intuition that such a peripheral change would in fact *not* affect one's experience, if central processing functioned as before.

we have for supposing a difference in experience will be entirely derivative on our reasons for supposing a difference in central state. Such a case therefore cannot hurt the invariance thesis.

Again, this argument might count against a coarse-grained or behavioral invariance thesis, according to which qualia are dependent only on discriminative capacities. The argument could also refute a view on which any two systems with different functional organization have different qualia, but few have advocated such a view. The invariance thesis according to which qualia are determined by fine-grained functional organization is unaffected.

7.5 Dancing Qualia[13]

One might think that the Fading Qualia argument could be directly adapted to provide an argument against the possibility of inverted qualia. Unfortunately this will not work. Imagine how an analogous argument would go. We start with me, having a red experience, and an inverted system having a blue experience. By gradual replacement, we construct a series of cases, each having some intermediate color. But there is nothing wrong with this! The intermediate systems are simply cases of mild qualia inversion, and are no more problematic than the extreme case.

To be sure, it may not be obvious just what the intermediate systems are experiencing. Perhaps no color from our usual color space can do the job, consistently with the systems' patterns of categorization and differences. But perhaps they are experiencing entirely new colors, ones that I cannot experience but that nevertheless form a continuum from red to blue. This would be odd, but it is not vastly implausible. Importantly, the problem with the Fading Qualia case will be entirely absent. These systems will *not* be systematically wrong about the features of their experience. Where they claim to experience distinctions, they may still experience distinctions; where they claim intense experiences, they have intense experiences; and so on. To be sure, the colors they call "red" will be different from what I call red, but this is nothing problematic; it happens already in the usual inversion case. What counts is that unlike the Fading Qualia case, the *structural* features of these systems' experiences are preserved throughout.

There is nevertheless a good argument against the possibility of inverted qualia to be found in this vicinity. Once again, for the purposes of *reductio*, assume that

[13]The argument in this section is distantly inspired by Dennett's story "Where Am I?" (Dennett 1978e). A situation bearing a certain resemblance to the one I describe below is considered by Shoemaker (1982). Perhaps the most closely related discussion can be found in Seager (1991, p.43), although Seager does not advocate the functional-invariance thesis.

inverted qualia are empirically possible. Then there can be two functionally isomorphic systems, in the same functional state but having different experiences. For the purposes of illustration let these systems be me experiencing red and my silicon isomorph experiencing blue, although as before the argument is general (with a caveat to be discussed).

As before, we construct a series of cases intermediate between me and my silicon isomorph. Here, the argument takes a different turn. We need not make any assumptions about the *way* in which experiences change as we move along the series. Perhaps they change suddenly, perhaps they jump all over the map, although surely it is most plausible that they change gradually. All that matters is that there must be two points A and B in this continuum of cases, such that (1) no more than 10% of the brain is replaced between A and B, and (2) A and B have significantly different experiences. This will surely be the case: we need only consider the points at which 10%, 20%, and so on up to 90% of the brain has been replaced. Red and blue are sufficiently different experiences that some neighboring pairs here *must* be significantly different (that is, different enough that the difference would be noticeable if they were experienced by the same person).

It is true that there can be unnoticeable differences between different experiences. If one changes a shade of red little enough, I will not be able to tell the difference. One might suppose that this is because there is no difference in experience, only a difference in the world; but if this were always the case one could iterate such a change 1000 times, eventually showing that red and blue produce the same experiences, which is ridiculous. So there can be *some* difference in experience that is not noticeable. One can observe this phenomenon by looking at a wide expanse of paint of subtly varying shade; sometimes it is extremely difficult to tell whether one's experiences of different parts is the same or different. But importantly, unnoticeable differences are very *small*. There is no way that ten unnoticeable jumps could take us all the way from red to blue. (It will be observed that this opens up a small loophole in the argument: what if the original inversion is only between two very similar experiences? I will return to this point.)

There must therefore be two systems A and B differing in at most 10% of their internal makeup, but having different experiences. For the purposes of illustration, let these systems be me and Bill. Where I have a red experience, Bill has a slightly different experience. We may as well suppose that Bill sees blue; perhaps his experience will be more similar to mine than that, but it makes no difference. Bill also differs in that where there are neurons in some small region of my brain, there are silicon chips in his brain. This substitution of a silicon circuit for a neural circuit is the only physical difference between Bill and me.

The crucial step in the thought-experiment is to take a silicon circuit just like

Bill's and install it in my own head as a *backup circuit*. This circuit will be functionally isomorphic to a circuit already present in my head. We equip the circuit with transducers and effectors so that it can interact with the rest of my brain, but we do not hook it up directly. Instead, we install a *switch* that can switch directly between the neural and silicon circuits. Upon flipping the switch, the neural circuit becomes irrelevant and the silicon circuit takes over. We can imagine that the switch controls the points of interface where the relevant circuits affects the rest of the brain. When it is switched, the connections from the neural circuit are pushed out of the way, and the silicon circuit's effectors are attached. (We can imagine that the transducers for both circuits are attached the entire time, so that the state of both circuits evolves appropriately, but so that only one circuit at a time is involved in processing. We can also run a similar experiment where both transducers and effectors are disconnected, so that the backup circuit is entirely isolated from the rest of the system. This would change a few details, but the moral would be the same.)

Immediately after the change, processing that was once performed by the neural circuit is now performed by the silicon circuit. One might say that the flow of control has been redirected. My functional organization remains the same throughout, however. All that changes is the physical makeup of one circuit, and the makeup of another circuit that is not even involved in processing. The uninvolved circuit is not part of my functional organization, as it plays no role in the directing of behavior.

What happens to my experience when we flip the switch? Before installing the circuit, I was seeing red. After we install it but before we flip the switch, I will presumably still be seeing red, as the only difference is the addition of a circuit that is not involved in processing in any way (I might as well have eaten the circuit for all the relevance it has to my processing). *After* flipping the switch, however, I am more or less the same system as Bill. The only difference is that there is a causally irrelevant neural circuit flapping in the breeze. (If someone objects that the neural circuit might make a difference afterwards just by being there, we can imagine destroying it as we make the change.) Bill, by hypothesis, was enjoying a blue experience. After the flip, then, I will see blue.

What will happen, then, is that my experience will change "before my eyes". Where I was once experiencing red, I will now experience blue. All of a sudden, I will have a *blue* experience of the apple on my desk. We can even imagine flipping the switch back and forth a number of times, so that the red and blue experiences "dance" before my eyes.

This might seem reasonable—it is a strangely appealing image—but something very odd is going on here. My experiences are switching from red to blue, but *I do not notice any change*. Even as we flip the switch a number of times and my qualia dance back and forth, I will simply go about my business, not noticing anything unusual. For my functional organization remains normal throughout. In particular, my functional organization after flipping the switch evolves just as it would have if the switch had not been flipped. There is no special difference in my behavioral dispositions. I am not suddenly disposed to say "Hmm! Something strange is going on!". There is no room for a sudden start, for an exclamation, or even for a distraction of attention. My cognitive organization is just as it usually is, and in particular is precisely as it would have been had the switch not been flipped.

Certainly, on any functional construal of belief, it is clear that I cannot acquire any new beliefs as the flip takes place. Even to one who disputes a functional account, it is is extremely implausible that a simple replacement of a neural circuit by a silicon circuit while overall organization is preserved could be responsible for the addition of significant new beliefs such as "My qualia just flipped". As in the case of Fading Qualia, there is simply no room for such a change to take place, unless it is in an accompanying Cartesian disembodied mind.

We are therefore led once more into a *reductio ad absurdum*. It seems entirely unreasonable to suppose that my experiences could change in such a significant way, even with me paying full attention, without my being able to notice the change. It would suggest once again a radical dissociation between consciousness and cognition. If this kind of thing could happen, then psychology and phenomenology would be quite out of step.

This sort of thing may be logically possible (although such a case is so extreme that it seems *only just* logically possible), but that does not mean we should take it seriously as an empirical possibility, any more than we should take seriously the possibility that the world was created five minutes ago. As an empirical hypothesis, it seems far more plausible that when one's experiences change significantly, then as long as one is rational and paying attention, one should be able to notice the change. If not, then consciousness and cognition are tied together only by the most slender of threads.

Indeed, if we are to suppose that Dancing Qualia like this are empirically possible, we are led to a worrying thought: they might be *actual*, and happening to us all the time. The physiological properties of our functional mechanisms are constantly changing. The functional properties of the mechanisms are reasonably robust; one would expect that this robustness would be ensured by evolution. But there is no adaptive reason for the non-functional properties to stay constant. From moment to moment there will certainly be changes in low-level molecular properties. Properties such as position, atomic makeup, and so on can take place while functional role is preserved, and are almost certainly going on constantly.

On the hypothesis that qualia are dependent not just on functional organization but on implementational details, it may well be that *our* qualia are in fact dancing before our eyes all the time. There seems to be no principled reason why a change from neurons to silicon should make a difference while a change in neural realization should not; the only place to draw a *principled* line is at the functional level[14] (White 1986 makes this sort of point, suggesting that if non-functional physical differences are relevant to qualia, then even tiny differences in DNA might affect qualia). The main reason for doubting that such dancing could be taking place is our belief in the following principle: when one's experiences change significantly, then one can notice the change. To accept the possibility of Dancing Qualia in the above case is to discard this principle, so it is no longer available as a defense against skepticism in our own case.

It is not entirely out of the question that we could actually *perform* such an experiment. Of course the practical difficulties would be immense, but at least in principle, one could install such a circuit in me and *I* could see what happened, and report it to the world. But of course there is no point performing the experiment: we know what the result will be. I will report that my experience stayed the same throughout, a constant shade of red, and that I noticed nothing untoward. I will become even more convinced than I was before that qualia are determined by functional organization. Of course this will not be a *proof*, but the evidence will nevertheless be hard to seriously dispute.

I conclude that by far the most plausible hypothesis is that replacement of neurons while preserving functional organization will preserve qualia, and that experience is wholly determined by functional organization.

Once again, one can extend the thought-experiment to other functional isomorphs. For systems much larger than a brain, we may need a complex system of radio transmitters to act as a connection between neurons and an external circuit, but that is no problem in principle. A problem arises with isomorphs that are much faster or slower than the original system. In this case, we cannot simply substitute a circuit from one system into the other and expect everything to function normally. However, we can still perform the experiment on a slowed-down or speeded-up version of the system in question. At worst, we have left open the possibility that a change in speed might invert qualia; but this hypothesis was never very plausible in the first place.

It should be noted that this thought-experiment works just as well against the possibility of absent qualia as against that of inverted qualia. We simply take two points on the way to absent qualia between which experience differs significantly, and install a backup circuit in the same way. As before, if absent qualia are possible, then switching will cause my qualia to oscillate before my eyes, from vivid to tepid and back, without my ever noticing. Again, it is far more plausible that such dancing without noticing is impossible, so that absent qualia are impossible.

[14]Shoemaker (1982) gives a complex criterion for how specific a physiological property needs to be to fix qualia, or to "realize a quale", as he puts it. However, it seems to me that if my discussion here has been correct, his criterion will in fact pick out a fine-grained functional property.

Personally, I find this an even more convincing argument against absent qualia than the argument in 7.3, although both have a role to play. An opponent might just bite the bullet and accept the possibility of Fading Qualia, but Dancing Qualia seem an order of magnitude more difficult to accept. The very immediacy of the switch seems to make a significant difference, as does the fact that the phenomenon the subject cannot notice is so dynamic and striking. Fading Qualia would mean that some systems are out of touch with their conscious experience, but Dancing Qualia would establish an even stranger gap.

I will respond to some objections. The objections to the Fading Qualia case can all be made again, and the replies are more or less the same; I will not bother to repeat them.

Objection 1: What about mild inversions?

As I noted earlier, this argument does not refute the possibility of very mild spectrum inversions—for instance, wherein two functionally isomorphic systems experience dark red and darker red respectively. In such a case there may be a sequence of nine intermediate shades with no noticeable difference between neighboring shades. On installing a circuit and flipping the switch, no change will be noticeable, but in this case that implies nothing unusual.

Of course, there is nothing special about the number 10. I picked this number because it allows that only a small amount (10%) of the brain needs to be replaced when we flip the switch. If it were too much more— say 50%—then we would start to have worries about personal identity. Is it *really* the same person experiencing the new color? If not, then the failure to notice a change would not be such a problem. Still, we can probably go up to 25% or 33% without a problem. This still leaves cases where two experiences are distinguishable without there being any intermediate experience distinguishable from both, however.

One can also reduce the impact of this problem by noting that it is unlikely that experience depends uniformly on all areas of the brain. It is likely that visual experience, say, is not dependent at all on the state of motor areas, auditory areas, and so on, or at least that it is dependent on these only insofar as they affect other areas. Perhaps it is dependent only on a small area of the visual cortex and central systems. If so, then it is possible that we could replace the entire area responsible for the experience in one fell swoop, while still only replacing a small amount of the brain. If we could do this, then the argument would still go through even for the mildest noticeable change experience.

Perhaps the best response, though, is simply to note that the possibility of mild underdetermination of experience by organization is a very unthreatening one. If we like, we could easily accept it, noting that any differences between isomorphs would be so slight as to be uninteresting. More likely, we can note that this would seem an odd and unlikely way for the world to be. It would seem reasonable that experiences should be invertible across the board, or not invertible at all, but why should the world be such that a small inversion is possible but nothing more? This would seem quite arbitrary. We cannot rule it out, but it is not something that we need to take seriously.

Objection 2: Double switching

Another objection (due to Terry Horgan) is the following. We can imagine a related experiment in which we rewire the connections from red and blue inputs to central areas of the brain so that blue inputs play the role that red inputs once played, and in which we also systematically rewire connections *downstream* from the central area to compensate. When a blue input causes the central area to go into a state previously associated with red, connections from the central area to the rest of the brain are rewired so that the rest of the brain functions just as it would have had there been no rewiring at all. In this way, my experience will almost certainly switch from red to blue, but my behavioral dispositions will stay constant throughout. In this case, a repeated switch would surely lead to Dancing Qualia. So aren't Dancing Qualia reasonable after all?

First, I should note that this rewiring would be a much vaster task than any other cases I have described. The central area will affect the rest of the brain at all sorts of different places. Each of these connections will have to be rewired, and crucially, no simple rewiring could do the job at any of them. We cannot simply switch "red outputs" to "blue outputs", as we could with inputs; the outputs from the central system may represent such diverse things as retrieved memories, motor instructions, and so on, with no simple difference in "polarity" between an output for red and an output for blue. To determine an appropriate "blue output", one would probably need to simulate the entire processing of the central area, given its initial state and input, to see what it produces. If so, it will be this simulation that is doing the causal work, not the central area itself, and the force of the scenario will be lost.

Second, even if there somehow turned out to be a simple way in which outputs could be rewired, note that *only* behavioral dispositions are preserved, and not functional organization. What might this feel like? In this case, I imagine that I would *notice* the switch and try to act accordingly, but would feel as if some jarring puppeteer was interfering with my actions. Unlike the previous case, there will be *room* for these extra beliefs and other cognitive states; they will be supported by the different states of the central area. And we can imagine that once feedback takes place, and input to the central areas indicates that its motor movements have been entirely different from what was planned, we can imagine that the central area state will be severely shaken up. In fact, this leads us back to the first objection, as it would seem almost impossible to systematically compensate for these feedback effects. In any case, the significant difference in functional organization means that the cases are not analogous.

7.6 Where things stand

To broadly summarize what has gone before: we have established that if absent qualia are possible, then Fading Qualia are possible; if inverted qualia are possible, then Dancing Qualia are possible; and as a bonus, if absent qualia are possible, then Dancing Qualia are possible. But it is very implausible that Fading Qualia and Dancing Qualia are possible. Therefore absent qualia and inverted qualia are impossible. The invariance thesis is true, and functional organization fully determines conscious experience.

It should be noted that we have only established the empirical or nomic impossibility of absent and inverted qualia. These arguments cannot be extended into an argument for their logical or metaphysical impossibility, as some functionalists might like. The first reason for this is that both Fading Qualia and Dancing Qualia are coherent hypotheses, if very implausible. Some might dispute the logical possibility of these hypotheses, holding perhaps that it is a constitutive property of qualia that we can notice differences in them (Shoemaker holds a position like this, I believe), but in any case there is a second, stronger reason why these arguments do not establish logical impossibility.

This second reason lies in the fact that both arguments have an empirical premise: that *I* have conscious experience, or that some biological system like me could have conscious experience. *Given* that I have conscious experience, then if Fading and Dancing Qualia are impossible, absent and inverted qualia are impossible, but the conclusion will be only as strong as the premise. It is an *empirical* fact, at best a nomic necessity, that a system like me will have conscious experience. It follows that it is at best nomically necessary that my functional isomorph will have the same sort of conscious experience.

Of course, if one could independently establish that the existence and nature of my experience is logically or metaphysically entailed by my physical structure, then this could be exploited (in combination with the logical impossibility of Fading and Dancing Qualia) to establish the logical impossibility of absent qualia and inverted qualia. But one could not establish the premise, as I have argued. If one *could* establish the premise, it would almost certainly be through some functional definition or analysis of qualia, and in this case the logical possibility of absent or inverted qualia would follow immediately without the need for complex arguments.

It follows that either way my arguments do not give the *reductive* functionalist position any particular support. One cannot *identify* qualia with functional states. One can, however, note that they are nomically *determined* by functional states.

We have therefore made a significant advance in our quest to constrain the principles in virtue of which consciousness nomically supervenes on the physical. We have narrowed down the relevant properties in the supervenience base to properties of *functional organization*. (The notion of functional organization is formalized in Chalmers (1993a) and will be discussed in a ninth chapter, summarized in the appendix, but the informal understanding has sufficed for our purposes here.)

In a certain sense, we can say that not only does consciousness supervene on the physical, but it supervenes on the functional. This needs to be spelled out carefully, due to the fact that every system realizes numerous kinds of functional organization, but we can say the following: for every physical system that gives rise to conscious experience, there is some functional organization F realized by the system, such that it is nomically necessary that any system that realizes F will have identical conscious experiences.

This alone does not say how we pick out the relevant F, but we can do this by ensuring that we go to a fine enough grain that F fixes our cognitive states, such as beliefs. This in turn we can ensure by going to a low enough level that F fixes the mechanisms responsible for the production of behavior, and fixes our behavioral dispositions. This is all that the Fading and Dancing Qualia arguments required, so it is all we need for functional invariance.

It is therefore a law, for certain functional organizations F, that realization of F will be accompanied by such-and-such conscious experience. This is not to say that it need be a *fundamental* law. It would be odd if the universe had fundamental laws connecting complex functional organizations to conscious experiences. Rather, one would expect that it would be a consequence of some simpler fundamental psychophysical laws. These fundamental laws remain to be determined, although I move some distance in that direction in work described in the appendix.

The functional-invariance thesis has significant consequences, of course. Not only does it give us a handle on some aspects of the problem of other minds; it tells us that in principle, cognitive systems realized in all sorts of other media can in principle be conscious. In particular the conclusion gives strong support to the ambitions of researchers in artificial intelligence, who hope that one day they might produce a conscious computational device. Artificial intelligence still faces many obstacles, but consciousness poses no insurmountable problem for it.

7.7 Overall conclusion

My project in this work has had two parts. First, I have argued for a non-reductive view of consciousness. Second, I have been concerned to establish that a non-reductive view does not rule out progress toward a theory of consciousness within a naturalistic framework. The overturning of the reductive view requires us to radically revise our picture of the natural order; it means that to have a hope of being successful, a theory of consciousness must be quite unlike a theory of most natural phenomena. Nevertheless, once overoptimistic hopes for a reductive explanation are out of the way, real progress toward a theory can begin.

The last few chapters have illustrated the way in which we might come to understand conscious experience naturalistically but non-reductively. In particular, we have come to a better understanding of the principles in virtue of which consciousness lawfully supervenes on the physical. The principle of structural coherence gave us a powerful handle on the relationship between consciousness and cognitive structure. The arguments in this chapter have provided good reason to believe that consciousness is a functionally invariant property, therefore strongly constraining the supervenience principles in question. Of course, this is just the beginning of what needs to be done, but it is a start.

Further considerations about the relationship between consciousness and phenomenal judgments, and about the relationship between consciousness and cognitive structure more generally, should allow us to constrain a theory of consciousness even further. Further arguments using thought-experiments might also play a role. Eventually we will be in a position to put forward fundamental theories of consciousness, and to evaluate them according to how well they meet these and other constraints. I give an example of such a theory in an appendix. Even if this theory is not the right theory, it does not matter too much. What matters is that we are moving toward a position where such theories can be put forward and evaluated. A complete understanding is still far off, but consciousness need not remain an eternal mystery.

Appendix: Future Work

There is much yet to say. Consciousness is a vast topic, and we have only scratched the surface. In particular, I have yet to put forward a full-fledged theory of consciousness that meets the constraints I have laid down. In future work, I hope to spell out a speculative theory that satisfies these constraints. I also hope that the conclusions I have reached about consciousness can be applied to central issues in artificial intelligence and in quantum mechanics. I will briefly summarize the possibilities for future work in each of these three directions below.

Consciousness and Information—A Theory Sketch

There are a number of constraints that we would like a theory of consciousness to satisfy. It should be compatible with the functional invariance thesis, so that experience arises from some aspect of functional organization. It should be compatible with—and ideally it should explain—the principle of structural coherence, according to which the contents of consciousness parallel the contents of awareness. It should further satisfy the requirement of explanatory coherence: the account of consciousness given by such a theory must bear an appropriate relation to an account of the the things we say and judge about consciousness. In this work, I will put such a theory forward. It is highly tentative, quite sketchy, and it may well be false, but at least it gives an idea of what a theory of consciousness might look like.

The theory is a double-aspect theory based on the notion of *information.* This is information not in the semantic sense of Dretske (1981), but in the syntactic/causal sense of Shannon (1948). Roughly, information is understood as a *difference that makes a difference* on some causal pathway. Information is best represented as point in a difference-manifold – a space of possible states that some part of a system can take on, individuated according to their effects on some causal pathway. Shannon's discussion centered on analyses of the *amount* of information involved in various causal transactions, where this amount is measured in bits. My discussion will be centered on information itself; we might say that information is to amount of information as matter is to mass. Essentially, I will reify the construct in order that it can serve as the basis of a fundamental theory. The theory holds, at a first approximation, that experience is another aspect of information—that is, information has both physical and phenomenal aspects with an irreducible connection between them. The dual-aspect nature of information is the fundamental nexus whereby consciousness arises from the physical.

This theory can be motivated in three ways. The first comes from considering the question: When does a physical difference give rise to a phenomenal difference? Some physical changes to a system make no difference to conscious experience. Changes in the wrong area, for instance, or changes to an irrelevant property (perhaps temperature) or an irrelevant part (perhaps a glial cell) will leave experience intact. The answer seems to be that a physical difference gives rise to a phenomenal difference when it makes a *causal* difference along certain pathways, particularly those responsible for verbal reports and other aspects of behavior. We note then that changes in experience correspond precisely to changes in information, supporting the theory.

Second, an information-based double-aspect theory is highly compatible with the principle of structural coherence. The structure of awareness can be represented directly as an informational state; differences in the structure correspond to differences in causal role down the same pathways mentioned above. Given the double-aspect nature of information, it follows that every difference in the structure of awareness will be mirrored by a difference in consciousness. With an appropriate account of the structure of information, the principle of structural coherence falls out.

Thirdly and importantly, the information-based theory meets the requirement of explanatory coherence perfectly. To demonstrate this, I will give a brief account of why we form the judgments that we do about consciousness. This is a reductive explanation, of course, and appeals to the fact that we have no access to our cognitive states over and above the information embodied therein. A cognitive system has no access to underlying neural properties, for instance, and no direct access to distal causes; it is simply presented with brute *differences* in cognitive state. It is thrust into different states without knowing why; if it is asked how one state is different from another, all it can say is "they're just different, *qualitatively*; this one is like *this*, and that one is like *that*." This leads naturally to judgments and reports about conscious experience. The basic explanation of our phenomenal judgments, then, lies in the fact that we only have access to the difference-structure of our cognitive states, or their informational properties. Information is at the explanatory basis for our phenomenal judgments, then, and by the principle of explanatory coherence we should expect to find it at the explanatory basis of consciousness itself. This is precisely what the theory suggests.

I will try to spell out the theory in some detail, although it will still be quite vague and sketchy. I will go on to address some very interesting questions that arise. Most obviously, there is information almost everywhere—in a simple nematode, even in a thermostat. Does this mean that those systems have *experience*? My answer to this is a tentative yes. Firstly it makes for by far the simplest theory: where there is information, there is consciousness. Secondly, it avoids having to posit any arbitrary cut-off points for conscious experience as we move down the scale of complexity. Third, it is compatible with a thesis about the coherence between sensation and perception. Of course the conclusion is counterintuitive at first, but I will argue that there is little ground for the intuition that simple systems cannot have experience, and that once we have accepted the irreducibility of experience, it is natural to expect that something so fundamental should be prevalent.

Wherever there is causation there is information. It follows that if this view is correct, wherever there is causation there is consciousness. This leads to some natural metaphysical speculation of the kind that we have canvassed already concerning a deep link between consciousness and causation; perhaps these two non-supervenient kinds are in some sense only one. I will develop this speculation further, and note the role that this theory might play in combating the epiphenomenalistic nature of nonreductive theories,

I will end with some grander metaphysical speculation: can we replace the dualism of the physical and the phenomenal with a monism of information? Some recent physical speculation has suggested that information may be in some sense the fundamental constituent of the universe, and that physical properties are derivative (this is the "it from bit" view; see Wheeler 1990). It is natural to suppose that the same might be true of consciousness, especially given our double-aspect theory. Developing such a monism of information is far from trivial and may ultimately be impossible, but it is an enticing goal.

Consciousness and Computation

Questions about consciousness are of central importance to artificial intelligence. Fundamental to the ambitions of artificial intelligence is a thesis of *computational sufficiency*, the view that Searle (1980) calls "Strong AI": that there exists a class of abstract computations (programs, algorithms, automata) such that implementation of any computation in that class suffices for mentality. For some aspects of mentality—those concerning behavioral capacities, for instance—the thesis is only somewhat controversial, but for consciousness, it is widely disputed.

Those who hold non-reductive and dualist views about consciousness are often disbelievers in the strong AI thesis, but the two issues are independent. I will argue that the strong AI thesis is correct, and that implementing the right computation suffices for consciousness.

To do this we first need an account of what it is to implement a computation. Searle (1990) and Putnam (1987) have argued that implementation conditions with the appropriate grade of specificity cannot be given, but such fears are ungrounded. I will introduce the notion of a *combinatorial state automaton* (CSA), a very general abstract computational object, and give an account of the conditions under which a given CSA is implemented. This provides a bridge between the theory of computation and physical theory. Roughly, the idea is that a physical system implements a computation if the causal state-transitional structure of the system mirrors the formal state-transitional structure of the computation (see Chalmers 1993).

This account of computation and implementation combines naturally with the functional invariance thesis to support the strong AI thesis. According to this account, a system implements a given computation if and only if it has a certain causal organization; according to the invariance thesis, all systems with an appropriate causal organization are conscious. Under certain natural auxiliary assumptions, the strong AI thesis falls out. One can also derive the thesis by running the Fading Qualia and Dancing Qualia arguments directly on any implementation of an appropriate CSA implemented by a conscious system. A gradual replacement scenario makes it clear that if one system is conscious, both are.

I will apply all this to critique the "Chinese room" argument of Searle 1980. Although this was put forward as an argument about intentionality, I will argue that it is mostly naturally seen as an argument about consciousness. The view I have developed supports a version of the "Systems Reply", on which the entire system has certain experiences even if a homunculus in the room does not. One can argue explicitly for this conclusion by running a version of the Fading Qualia argument involving a continuum of cases from the Chinese room to a human brain. If the Chinese room implements the right program, then there will be real physical causal organization between various symbols on paper within it, corresponding precisely to the causal structure among neurons in the brain. It is this causal structure that gives rise to consciousness.

I will also address Searle's "syntax is not sufficient for semantics" argument against strong AI, as well as a few other difficulties concerning the relationship between consciousness and computation.

Consciousness and Quantum Mechanics

The greatest mystery in physical theory remains the problem of interpreting quantum mechanics. Quantum mechanics provides an excellent *recipe* for predicting the results of physical observations. The recipe consists of two parts: a linear part, stating that most of the time the state of a physical system evolves as a wavefunction according to Schrödinger's equation, and a nonlinear part, stating that whenever a measurement is made the wave "collapses" in a nonlinear, nondeterministic way, producing some kind of discrete state, according to probabilities determined by the wavefunction. This recipe gets observations precisely right, but it is very hard to see what might be really going on in the world in order that such a recipe should come out right.

One might think that one could take the theory at face value, holding that the state of a physical system is a superimposed wave that collapses upon measurement, but the problem lies in the notion of measurement. On the face of it, this is a vague high-level term that is quite inappropriate for a physical theory. The problem of quantum measurement is to find some non-arbitrary criterion for what counts as a *measurement* in order that the theory can give determinate conditions for the occurrence of a collapse, or to find some alternative interpretation of the theory that explains its predictive success.

A number of suggestions have been put forward here, but most have significant problems. First, it has sometimes been supposed that one can define a measurement as an interaction with a macroscopic object such as a measuring device; but the notion of a macroscopic object is vague, and in any case there seems no reason to suppose that macroscopic objects themselves should not be bound by quantum laws.

Second, it is sometimes suggested that a measurement is defined by the presence of *consciousness*, and that consciousness itself is responsible for collapsing the wavefunction. This is perhaps an odd view coming from physicist, as it requires the truth of dualism for it to be remotely tenable. Despite this compatibility with dualism, I do not advocate the view. It is incompatible with a view on which consciousness is ubiquitous (as I suggested earlier), as that would imply a constantly-collapsing wavefunction of a kind incompatible with the empirical data. More generally, it is very difficult to give an account of what *kind* of interaction with consciousness suffices to collapse the wavefunction, and the question of *how* the function collapses is entirely mysterious.

A third possibility is a "hidden-variables" view upon which there is no collapse, and on which the wavefunction is merely an incomplete description of a deeper underlying reality. Theories like this compatible with the empirical data can be devised, but they are clumsy and ad hoc, and furthermore require nonlocal interaction between parts of a physical system, contrary to a central tenet of modern physics.

Fourthly one can deny that there is any fact of the matter about a system that is not being observed. Bohr's "Copenhagen interpretation" is a version of this view, popular among physicists because of its dismissal of metaphysical difficulties, but ultimately it seems only to duck the question. Taken seriously, it requires embracing the phenomenalist view that the only reality derives from our experience, with all the problems attendant on such a view.

A fifth and interesting possibility holds that the world is fully described by Schrödinger's equation and that there is no "collapse". The trouble with this view is that it seems to predict that the world should be a superposition at the macroscopic level, with pointers pointing to many positions simultaneously, cats that are both dead and alive, and so on; and this seems to contradict the manifest facts. Some physicists have suggested that Schrödinger's equation might be compatible with discreteness at the macroscopic level, perhaps if wave evolution involved certain "amplification" effects, but this seems to be mathematically untenable. We are faced with a general problem: if the world is a superposition at the microscopic level, why does it seem discrete at the macroscopic level?

Perhaps the most radical response is due to Everett (1957). This holds that the world is fully described by Schrödinger's equation, and accepts that the world is a superposition even at the macroscopic level. Pointers point to many positions simultaneously, cats are both dead and alive, and so on. This may seem contrary to experience, but Everett notes that observers themselves are quantum systems that consist in superpositions of many states. It is possible that relative to such a state, the world could be perceived as discrete (Everett calls his view the "relative state" interpretation). Given any superposed observing system, there will be many such subsidiary states, each perceiving a different discrete aspect of the world (a different aspect of the wavefunction). *I* correspond to just one of those subsidiary states.

The Everett view is often misdescribed as the "many-worlds" interpretation of quantum mechanics, upon which the world is constantly splitting into many different worlds. This version of the view is extremely problematic; the "branching" event it postulates is even more mysterious than "collapse". The Everett view need not involve any branching of worlds, however. It requires only a constantly evolving wavefunction that is a superposition of many parts.

The real problem with the Everett view is to answer the question: why does the world look like *this*, given that it is like *that*? Why are there observers (or subsidiary states of observers) who experience a discrete world, given the worlds superposed nature? This, I will argue, is not a problem about physics but a problem about consciousness. The question comes down to: why, given physical states like *this*, are there experiences like *that*? This is in the domain of a theory of consciousness.

I will argue that the theory of consciousness that I have developed is perfectly suited to handle this question. In fact, it turns out that the theory *predicts* that in a world made up of this sort of superimposed wavefunctions, there will be experiences as of a discrete world. This conclusion can be arrived at by noting that any information that would be present in a "collapsed" wavefunction will also be present in the full pre-collapse wavefunction; the collapsed function is merely an eigenstate of the original function, after all. If our information-based theory predicts that the "collapsed" discrete world will support a given experience, then, it will also predict that the uncollapsed world will support it. We can see the same conclusion a different way by applying the functional invariance thesis, and noting that any organization implemented in a "collapsed" world will also be realized in a corresponding uncollapsed world.

It follows that our theory *predicts* that even if the world is uncollapsed, there will be experiences just as there are in a collapsed world. In fact, it predicts that for many subjects, the quantum recipe will predict the results of observations perfectly. The lack of collapse is therefore no problem for the Everett interpretation, if it is supplemented by an appropriate theory of consciousness. The world out there is superimposed, but parts of it give rise to experiences that are discrete.

I will argue that the Everett view provides the best interpretation of quantum mechanics. Its underlying physical principles are simpler than those of any remotely viable alternative: the world obeys the Schrödinger equation, and that is that. And it predicts that data perfectly, when supplemented with an independently-motivated theory of consciousness. As the simplest theory compatible with the data, we should accept it.

I will discuss various objections to the Everett view that have been put forward, and argue that none of these are persuasive. Many of these are based on misinterpreting it as requiring a literal branching of worlds; others underestimate the role that a good theory of consciousness can play in justifying the view. The view has some counterintuitive consequences about personal identity, but I will be argue that these views are not real problems if we adopt the view of personal identity advocated by Parfit (1984). Finally, the view is counterintuitive in its implication that there is far more in the world than we ever experience directly, and we may never be able to accept it emotionally, but we should at least take seriously the possibility that it is true.

Bibliography

Ackermann, D. 1990. *A Natural History of the Senses.* Random House.

Adams, R.M. 1974. Theories of actuality. *Nous* 8:211-31.

Akins, K. 1989. What is it like to be boring and myopic? Manuscript.

Alexander, S. 1920. *Space, Time, and Deity.* Macmillan.

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

Armstrong, D.M. 1973. *Belief, Truth, and Knowledge.* Cambridge University Press.

Armstrong, D.M. 1981. What is consciousness? In *The Nature of Mind.* Cornell University Press.

Armstrong, D.M. 1982. Metaphysics and supervenience. *Critica* 42:3-17.

Armstrong, D.M. 1983. *What is a Law of Nature?* Cambridge University Press.

Armstrong, D.M. 1990. *A Combinatorial Theory of Possibility.* Cambridge University Press.

Baars, B.J. 1988. *A cognitive theory of consciousness.* Cambridge University Press.

Bacon, J. 1986. Supervenience, necessary coextensions, and reducibility. *Philosophical Studies* 49:163-76.

Barwise, J. and Perry, J. 1983. *Situations and Attitudes.* MIT Press.

Bigelow, J. and Pargetter, R. 1990. Acquaintance with qualia. *Theoria.*

Bisiach, E. 1988. The (haunted) brain and consciousness. In (A. Marcel & E. Bisiach, eds) *Consciousness in Contemporary Science.* Oxford University Press.

Blackburn, S. 1971. Moral realism. In (J. Casey, ed.) *Morality and Moral Reasoning.* Methuen.

Blackburn, S. 1985. Supervenience revisited. In (I. Hacking, ed.) *Exercises in Analysis: Essays by Students of Casimir Lewy.* Cambridge University Press.

Block, N. 1978. Troubles with functionalism. In (C.W. Savage, ed.) *Perception and Cognition: Issues in the Foundation of Psychology.* University of Minnesota Press.

Block, N. 1980a. Troubles with functionalism. In (N. Block, ed.) *Readings in the Philosophy of Psychology*, Vol 1. Harvard University Press.

Block, N. 1980b. What is functionalism? In (N. Block, ed.) *Readings in the Philosophy of Psychology*, Volume 1. Harvard University Press.

Block, N. 1981. Psychologism and behaviorism. *Philosophical Review* 90:5-43.

Block, N. 1986. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy* 10:615-78.

Block, N. 1990a. Consciousness and accessibility. *Behavioral and Brain Sciences* 13:596-98.

Block, N. 1990b. Inverted earth. *Philosophical Perspectives* 4:53-79.

Block, N. 1993. Review of D.C. Dennett, *Consciousness Explained*. *Journal of Philosophy* 90:181-93.

Bogen, J. 1981. Agony in the schools. *Canadian Journal of Philosophy* 11:1-21.

Boyd, R. 1980. Materialism without reductionism: What physicalism does not entail. In (N. Block, ed.) *Readings in the Philosophy of Psychology*, Volume 1. Harvard University Press.

Boyd, R.N. 1988. How to be a moral realist. In (G. Sayre-McCord, ed.) *Essays on Moral Realism*. Cornell University Press.

Brink, D. 1989. *Moral realism and the foundations of ethics*. Cambridge University Press.

Burge, T. 1979. Individualism and the mental. *Midwest Studies in Philosophy* 4:73-122.

Burge, T. 1986. Individualism and psychology. *Philosophical Review* 95:3-45.

Campbell, C. 1970. *Body and Mind*. Doubleday.

Carroll, J. 1990. The Humean tradition. *Philosophical Review* 99:185-219.

Chalmers, D.J. 1993a. A computational foundation for the study of cognition. *Minds and Machines*, forthcoming.

Chalmers, D.J. 1993b. Does a rock implement every finite state automaton? Manuscript.

Cheney, D.L., and Seyfarth, R.M. 1990. *How Monkeys See the World*. University of Chicago Press.

Childs, W. 1993. Anomalism, uncodifiability, and psychophysical relations. *Philosophical Review* 102.

Chisholm, R. 1957. *Perceiving.* Cornell University Press.

Churchland, P.M. 1981. Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78:67-90.

Churchland, P.M. 1985. Reduction, qualia and the direct introspection of brain states. *Journal of Philosophy* 82:8-28.

Churchland, P.M. and Churchland, P.S. 1981. Functionalism, qualia and intentionality. *Philosophical Topics* 12:121-32.

Churchland, P.S. 1988. The significance of neuroscience for philosophy. *Trends in the Neurosciences* 11:304-307.

Clark, A. 1986. Psychofunctionalism and chauvinism. *Philosophy of Science* 53:535-59.

Clark, A. 1989. *Microcognition.* MIT Press.

Cole, D. 1990. Functionalism and inverted spectra. *Synthese* 82:207-22.

Crick, F. and Koch, C. 1990. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2: 263-275.

Cuda, T. 1985. Against neural chauvinism. *Philosophical Studies* 48:111-27.

Davidson, D. 1970. Mental events. In (L. Foster and J. Swanson, eds.) *Experience and Theory.* Humanities Press.

Davies, M.K. and Humberstone, I.L. 1980. Two notions of necessity. *Philosophical Studies* 38:1-30.

Dennett, D.C. 1968. *Content and Consciousness.* Routledge and Kegan Paul.

Dennett, D.C. 1978a. *Brainstorms.* MIT Press.

Dennett, D.C. 1978b. Why you can't make a computer that feels pain. In Dennett 1978a.

Dennett, D.C. 1978c. Toward a cognitive theory of consciousness. In Dennett 1978a.

Dennett, D.C. 1978d. Are dreams experiences? In Dennett 1978a.

Dennett, D.C. 1978e. Where am I? In Dennett 1978a. MIT Press.

Dennett, D.C. 1979. On the absence of phenomenology. In (D. Gustafson and B. Tapscott, eds) *Body, Mind, and Method.* Kluwer.

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

Dennett, D.C. 1988. Quining qualia. In (A. Marcel and E. Bisiach, eds.) *Consciousness in Contemporary Science.* Oxford University Press.

Dennett, D.C. 1991. *Consciousness Explained.* Little-Brown.

Dretske, F.I. 1970. *Seeing and Knowing.*

Dretske, F.I. 1977. *Laws of Nature.* Philosophy of Science 44:248-68.

Dretske, F.I. 1981. *Knowledge and the Flow of Information.* MIT Press.

Eccles, J.C. 1986. Do mental events cause neural events analogously to the probability fields of quantum mechanics? *Proceedings of the Royal Society of London B* 227:411-28.

Edelman, G. 1989. *The Remembered Present: A Biological Theory of Consciousness.* Basic Books.

Edelman, G. 1992. *Bright Air, Brilliant Fire.* Basic Books.

Elitzur, A. 1989. Consciousness and the incompleteness of the physical explanation of behavior. *Journal of Mind and Behavior* 10:1-20.

Everett, H. 1957. 'Relative-state' formulations of quantum mechanics. *Reviews of Modern Physics* 29: 454-62.

Feldman, F. 1974. Kripke on the identity theory. *Journal of Philosophy* 71:665-76.

Flanagan, O. 1992. *Consciousness Reconsidered.* MIT Press.

Fodor, J.A. 1975. *The Language of Thought.* Harvard University Press.

Fodor, J.A. 1987. *Psychosemantics.* MIT Press.

Fodor, J.A. 1992. The big idea: Can there be a science of mind? *Times Literary Supplement* 4567:5-7 (July 3 1992).

Forrest, P. 1986. Ways worlds could be. *Australasian Journal of Philosophy* 64:15-24.

Foss, J. 1989. On the logic of what it is like to be a conscious subject. *Australasian Journal of Philosophy* 67:305-320.

Geach, P. 1957. *Mental Acts.* Routledge and Kegan Paul.

Gert, B. 1965. Imagination and verifiability. *Philosophical Studies* 16:44-47.

Goldman, A. 1986. *Epistemology and Cognition.* Harvard University Press.

Goldman, A. 1993. The psychology of folk psychology. *Behavioral and Brain Sciences.*

Grossman, R. 1992. Materialism and the new folk philosophy. Manuscript.

Hardin, C.L. 1987. Qualia and materialism: Closing the explanatory gap. *Philosophy and Phenomenological Research* 48:281-98.

Hardin, C.L. 1988. *Color for Philosophers.* Hackett.

Hare, R.M. 1952. *The Language of Morals.* London.

Harman, G. 1982. Conceptual role semantics. *Notre Dame Journal of Formal Logic* 28:242-56.

Harrison, B. 1967. On describing colors. *Inquiry* 10:38-52.

Harrison, B. 1973. *Form and Content.* Blackwell.

Harrison, J. 1981a. Three philosophical fairy stories. *Ratio* 23:63-67.

Harrison, J. 1981b. Gulliver's adventures in Fairyland. *Ratio* 23:158-64.

Haugeland, J. 1982. Weak supervenience. *American Philosophical Quarterly* 19:93-103.

Heil, J. 1992. *The Nature of True Minds.* Cambridge University Press.

Heil, J., and Mele. A. (eds.) 1992. *Mental Causation.* Oxford University Press.

Hellman, G. and Thompson, F. 1975. Physicalism: ontology, determination and reduction. *Journal of Philosophy* 72:551-64.

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

Hofstadter, D.R. 1979. *Gödel, Escher, Bach: an Eternal Golden Braid.* Basic Books.

Hofstadter, D.R. 1985. Who shoves whom around inside the careenium? In *Metamagical Themas.* Basic Books

Honderich, T. 1981. Psychophysical law-like connections and their problems. *Inquiry* 24:277-303.

Horgan, T. 1978. Supervenient bridge laws. *Philosophy of Science* 45: 227-49.

Horgan, T. 1982. Supervenience and microphysics. *Pacific Philosophical Quarterly* 63: 29-43.

Horgan, T. 1984a. Functionalism, qualia, and the inverted spectrum. *Philosophy and*

Phenomenological Research 44:453-69.

Horgan, T. 1984b. Jackson on physical information and qualia. *Philosophical Quarterly* 34:147-83.

Horgan, T. 1984c. Supervenience and cosmic hermeneutics. *Southern Journal of Philosophy Supplement* 22:19-38.

Horgan, T. 1987. Supervenient qualia. *Philosophical Review* 96: 491-520.

Horgan, T., and Timmons, M. 1992a. Troubles for new wave moral semantics: The "open question argument" revived. *Philosophical Papers.*

Horgan, T., and Timmons, M. 1992b. Trouble on moral twin earth: Moral queerness revived. *Synthese.*

Huxley, T. 1874. On the hypothesis that animals are automata. In *Collected Essays.* London, 1893-4.

Jackendoff, R. 1987. *Consciousness and the Computational Mind.* MIT Press.

Jackson, F. 1977. *Perception.* Cambridge University Press.

Jackson, F. 1980. A note on physicalism and heat. *Australasian Journal of Philosophy* 58:26-34.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32: 127-36.

Jackson, F. 1993. Armchair metaphysics. In (J. O'Leary-Hawthorne and M. Michael, eds.) *Philosophy in Mind.* Kluwer.

Jacoby, H. 1990. Empirical functionalism and conceivability arguments. *Philosophical Psychology* 2:271-82.

Jaynes, J. 1976. *The Origins of Consciousness in the Breakdown of the Bicameral Mind.* Houghton Mifflin.

Johnson-Laird, P. 1983. A computational analysis of consciousness. *Cognition and Brain Theory* 6:499-508.

Kaplan, D. 1978. On the logic of demonstratives. *Journal of Philosophical Logic* 8:81-98.

Kaplan, D. 1979. *Dthat.* In (P. Cole, ed.) *Syntax and Semantics.* New York: Academic Press.

Kaplan, D. 1989. Demonstratives. In (J. Almog, J. Perry, and H. Wettstein, ed.) *Themes from Kaplan.* Oxford University Press.

Kim, J. 1978. Supervenience and nomological incommensurables. *American Philosophical Quarterly* 15:149-56.

Kim, J. 1984. Concepts of supervenience. *Philosophy and Phenomenological Research* 45:153-76.

Kim, J. 1985. Psychophysical laws. In (B. McLaughlin & E. LePore, eds) *Action and Events*. Blackwell.

Kim, J. 1989. Mechanism, purpose, and explanatory exclusion. *Philosophical Perspectives* 3:77-108.

Kirk, R. 1974. Zombies versus materialists. *Aristotelian Society Supplement* 48: 135-52.

Kirk, R. 1979. From physical explicability to full-blooded materialism. *Philosophical Quarterly* 29:229-37.

Kripke, S. 1972. Naming and necessity. In (G. Harman and D. Davidson, eds.) *The Semantics of Natural Language*, p. 254-355. Dordrecht. Reprinted as *Naming and Necessity* (1980). Harvard University Press

Lahav, R. and Shanks, N. 1992. How to be a scientifically respectable "property dualist". *Journal of Mind and Behavior* 13:211-32.

Langton, C.G. 1989. *Artificial Life: The proceedings of an interdisciplinary workshop on the synthesis and simulation of living systems.* Addison-Wesley.

Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-61.

Levine, J. 1988. Absent and inverted qualia revisited. *Mind and Language* 3:271-87.

Levine, J. 1991. Cool red. *Philosophical Psychology* 4:27-40.

Lewis, D. 1966. An argument for the identity theory. *Journal of Philosophy* 63:17-25.

Lewis, D. 1973. *Counterfactuals.* Harvard University Press.

Lewis, D. 1974. Radical interpretation. *Synthese* 23:331-44.

Lewis, D. 1977. Attitudes *de dicto* and *de se*. *Philosophical Review* 88:513-43.

Lewis, D. 1978. Mad pain and martian pain. In (N. Block, ed.) *Readings in the Philosophy of Psychology*, Vol. 1. MIT Press.

Lewis, D. 1979. Attitudes *de dicto* and *de se*. *Philosophical Review* 88:513-45.

Lewis, D. 1983. New work for a theory of universals. *Australasian Journal of Philosophy.*

Lewis, D. 1984. Putnam's paradox. *Australasian Journal of Philosophy* 62:221-36.

Lewis, D. 1986a. *On the Plurality of Worlds.* Blackwell.

Lewis, D. 1986b. *Philosophical Papers, Volume II.* Oxford University Press.

Lewis, D. 1990. What experience teaches. In (W. Lycan, ed.) *Mind and Cognition.* Blackwell.

Lewis, D. 1994. Reduction of mind. In (S. Guttenplan, ed.) *A Companion to the Philosophy of Mind.* Blackwell.

Loar, B. 1988. Social content and psychological content. In (R. Grimm and D. Merrill, eds.) *Contents of Thought.* University of Arizona Press.

Loar, B. 1990. Phenomenal states. *Philosophical Perspectives* 4:81-108.

Lycan, W.G. 1973. Inverted spectrum. *Ratio* 15:315-9.

Lycan, W.G. 1987. *Consciousness.* MIT Press.

Mackie, J.L. 1974. *The Cement of the Universe.* Oxford University Press.

Mackie, J.L. 1977. *Ethics: Inventing Right and Wrong.* Penguin Books.

McGinn, C. 1977. Anomalous monism and Kripke's Cartesian intuitions. *Analysis* 2:78-80.

McGinn, C. 1989. Can we solve the mind–body problem? *Mind* 98:349-66.

McLaughlin, B.P. 1992. The rise and fall of the British emergentists. In (A. Beckermann, H. Flohr, and J. Kim, eds.) *Emergence or Reduction?: Prospects for Nonreductive Physicalism.* De Gruyter.

McMullen, C. 1985. 'Knowing what it's like' and the essential indexical. *Philosophical Studies* 48:211-33.

Meehl, P.E., and Sellars, W. 1956. The concept of emergence. In (H. Feigl and M. Scriven, eds.) *Minnesota Studies in the Philosophy of Science, Volume 1.* University of Minnesota Press.

Millikan, R.G. 1986. Thoughts without laws: Cognitive science with content. *Philosophical Review* 95:47-80.

Molnar, G. 1969. Kneale's argument revisited. *Philosophical Review* 78:79-89.

Moore, G.E. 1922. *Philosophical Studies.* Routledge and Kegan Paul.

Nagel, T. 1970. Armstrong on the mind. *Philosophical Review* 79:394-403.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 4:435-50.

Nagel, T. 1986. *The View from Nowhere.* Oxford University Press.

Natsoulas, T. 1978. Consciousness. *American Psychologist* 33:906-14.

Nelkin, N. 1989. Unconscious sensations. *Philosophical Psychology* 2:129-41.

Nemirow, L. 1990. Physicalism and the cognitive role of acquaintance. In (W. Lycan, ed.) *Mind and Cognition.* Blackwell.

Newell, A. 1992. SOAR as a unified theory of cognition: Issues and explanations. *Behavioral and Brain Sciences* 15:464-92.

Parfit, D. 1984. *Reasons and Persons.* Oxford University Press.

Penrose, R. 1989. *The Emperor's New Mind.* Oxford University Press.

Perry, J. 1977. Frege on demonstratives. *Philosophical Review* 86:474-97.

Perry, J. 1979. The problem of the essential indexical. *Nous* 13:3-21.

Petrie, B. 1987. Global supervenience and reduction. *Philosophy and Phenomenological Research* 48:119-30.

Place, U.T. 1956. Is consciousness a brain process? *British Journal of Psychology* 47:44-50.

Plantinga, A. 1976. Actualism and possible worlds. *Theoria* 42:139-60.

Pour-El, M.B., and Richards, J.I. 1989. *Computability in Analysis and Physics.* Springer-Verlag.

Putnam, H. 1960. Minds and machines. In (S. Hook, ed) *Dimensions of Mind.* New York University Press.

Putnam, H. 1975. The meaning of "meaning". In (K. Gunderson, ed.) *Language, Mind, and Knowledge.* University of Minnesota Press.

Putnam, H. 1981. *Reason, Truth, and History.* Cambridge University Press.

Putnam, H. 1983. Possibility and necessity. In *Reality and Reason: Philosophical Papers Volume 3.* Cambridge University Press.

Putnam, H. 1987. *Representation and Reality.* MIT Press.

Pylyshyn, Z. 1980. The "causal power" of machines. *Behavioral and Brain Sciences* 3:442-44.

Quine, W.V. 1951. Two dogmas of empiricism. *Philosophical Review* 60: 20-43.

Quine, W.V. 1969. Propositional objects. In *Ontological Relativity and Other Essays.* Columbia University Press.

Reynolds, C. 1987. Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics* 21(4): 25-34.

Robinson, H. 1976. The mind–body problem in contemporary philosophy. *Zygon* 11:346-360.

Rosenthal, D.M. 1990. A theory of consciousness. Bielefeld Report.

Ryle, G. 1949. *The Concept of Mind.*

Savitt, S. 1982. Searle's demon and the brain simulator reply. *Behavioral and Brain Sciences* 5:342-3.

Sayre-McCord, G. 1988. Introduction: The many moral realisms. In (G. Sayre-McCord, ed.), *Essays on Moral Realism.* Cornell University Press.

Schlick, M. 1932. Positivism and Realism. *Erkenntnis* 3.

Seager, W.E. 1988. Weak supervenience and materialism. *Philosophy and Phenomenological Research* 48:697-709.

Seager, W.E. 1991. *Metaphysics of Consciousness.* Routledge.

Searle, J.R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3: 417-57.

Searle, J.R. 1990. Consciousness, explanatory inversion and cognitive science. *Behavioral and Brain Sciences* 13:585-642.

Searle, J.R. 1992. *The Rediscovery of the Mind.* MIT Press.

Sellars, W. 1978. Is consciousness physical? Third Carus Lecture, Harvard University.

Shallice, T. 1972. Dual functions of consciousness. *Psychological Review* 79:383-93.

Shallice, T. 1988. Information-processing models of consciousness: possibilities and problems. In (A. Marcel and E. Bisiach, eds.) *Consciousness in Contemporary Science.* Oxford University Press.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379-423.

Shoemaker, S. 1975a. Functionalism and qualia. *Philosophical Studies* 27:291-315.

Shoemaker, S. 1975b. Phenomenal similarity. *Critica* 7:3-37.

Shoemaker, S. 1981. Some varieties of functionalism. *Philosophical Topics* 12:93-119.

Shoemaker, S. 1982. The inverted spectrum. *Journal of Philosophy* 79:357-381.

Sidelle, A. 1989. *Necessity, essence, and individuation.* Cornell University Press.

Sidelle, A. 1992. Rigidity, ontology, and semantic structure. *Journal of Philosophy* 8:410-30.

Siewert, C. 1993. What Dennett can't imagine and why. *Inquiry.*

Skyrms, B. 1980. *Causal Necessity.* Yale University Press.

Smart, J.J.C. 1959. Sensations and brain processes. *Philosophical Review* 68:141-56.

Sperry, R.W. 1969. A modified concept of consciousness. *Psychological Review* 76:532-36.

Sperry, R.W. 1992. Turnabout on consciousness: A mentalist view. *Journal of Mind and Behavior* 13:259-80.

Stalnaker, R. 1976. Possible worlds. *Nous* 10:65-75.

Stalnaker, R. 1978. Assertion. In (P. Cole, ed.) *Syntax and Semantics: Pragmatics, Vol. 9.* Academic Press, 1978.

Sutherland, N.S. (ed.) 1989. *The International Dictionary of Psychology.* New York: Continuum.

Taylor, D. 1966. The incommunicability of content. *Mind* 75:527-41.

Teller, P. 1984. A poor man's guide to supervenience and determination. *Southern Journal of Philosophy Supplement* 22:137-162.

Tienson, J.L 1987. Brains are not conscious. *Philosophical Papers* 16:187-93.

Tooley, M. 1977. The nature of laws. *Canadian Journal of Philosophy* 7: 667-98.

Tye, M. (forthcoming). Blindsight, the absent qualia hypothesis, and the mystery of consciousness. In (C. Hookway, ed.) *Philosophy and the Cognitive Sciences.* Cambridge University Press.

Tye, M. 1986. The subjective qualities of experience. *Mind* 95:1-17.

van Gulick, R. 1988. A functionalist plea for self-consciousness. *Philosophical Review* 97:149-88.

van Gulick, R. 1993. Understanding the phenomenal mind: Are we all just armadillos? In (M. Davies and G. Humphreys, eds.) *Consciousness: A Mind and Language Reader.* Blackwell.

Velmans, M. 1991. Is human information-processing conscious? *Behavioral and Brain Sciences* 14:651-69.

Wheeler, J.A. 1990. Information, physics, quantum: The search for links. In (W. Zurek, ed) *Complexity, Entropy, and the Physics of Information.* Addison-Wesley.

White, S.L. 1982. Partial character and the language of thought. *Pacific Philosophical Quarterly* 63:347-65.

White, S.L. 1986. Curse of the qualia. *Synthese* 68:333-68.

Winograd, T. 1972. *Understanding Natural Language.* Academic Press.

Wittgenstein, L. Notes for lectures on "private experience" and "sense data". *Philosophical Review* 77.

Wittgenstein, L. 1953. *Philosophical Investigations.* Macmillan.

Yablo, S. 1993. Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 53:1-42.