Facing Up to the Problem of Consciousness

David J. Chalmers
Department of Philosophy
Washington University
St. Louis, Missouri 63130
dave@twinearth.wustl.edu

December 16, 1994

1 Introduction

Consciousness poses the most baffling problems in the science of the mind. There is nothing that we know more intimately than conscious experience, but there is nothing that is harder to explain. All sorts of mental phenomena have yielded to scientific investigation in recent years, but consciousness has stubbornly resisted. Many have tried to explain it, but the explanations always seem to fall short of the target. Some have been led to suppose that the problem is intractable, and that no good explanation can be given.

To make progress on the problem of consciousness, we have to confront it directly. In this paper, I first isolate the truly hard part of the problem, separating it from more tractable parts and giving an account of why it is so difficult to explain. I critique some recent work that uses reductive methods to address consciousness, and argue that such methods inevitably fail to come to grips with the hardest part of the problem. Once this failure is recognized, the door to further progress is opened. In the second half of the paper, I argue that if we move to a new kind of nonreductive explanation, a naturalistic account of consciousness can be given. I put forward my own candidate for such an account: a nonreductive theory based on principles of structural coherence and organizational invariance, and a double-aspect theory of information.

2 The easy problems and the hard problem

There is not just one problem of consciousness. "Consciousness" is an ambiguous term, referring to many different phenomena. Each of these phenomena needs to be explained, but some are easier to explain than others. At the start, it is useful to divide the associated problems of consciousness into "hard" and "easy" problems. The easy problems of consciousness are those that seem directly susceptible to the standard methods of cognitive science, whereby a phenomenon is explained in terms of computational or neural mechanisms. The hard problems are those that seem to resist those methods.

The easy problems of consciousness include those of explaining the following phenomena:

- the ability to discriminate, categorize, and react to environmental stimuli;

- the integration of information by a cognitive system;
- the reportability of mental states;
- the ability of a system to access its own internal states;
- the focus of attention;
- the deliberate control of behavior;
- the difference between wakefulness and sleep.

All of these phenomena are associated with the notion of consciousness. For example, one sometimes says that a mental state is conscious when it is verbally reportable, or when it is internally accessible. Sometimes a system is said to be conscious of some information when it has the ability to react on the basis of that information, or, more strongly, when it attends to that information, or when it can integrate that information and exploit it in the sophisticated control of behavior. We sometimes say that an action is conscious precisely when it is deliberate. Often, we say that an organism is conscious as another way of saying that it is awake.

There is no real issue about whether *these* phenomena can be explained scientifically. All of them are straightforwardly vulnerable to explanation in terms of computational or neural mechanisms. To explain access and reportability, for example, we need only specify the mechanism by which information about internal states is retrieved and made available for verbal report. To explain the integration of information, we need only exhibit mechanisms by which information is brought together and exploited by later processes. For an account of sleep and wakefulness, an appropriate neurophysiological account of the processes responsible for organisms' contrasting behavior in those states will suffice. In each case, an appropriate cognitive or neurophysiological model can clearly do the explanatory work.

If these phenomena were all there was to consciousness, then consciousness would not be much of a problem. Although we do not yet have anything close to a complete explanation of these phenomena, we have a clear idea of how we might go about explaining them. This is why I call these problems the easy problems. Of course, "easy" is a relative term. Getting the details right will probably take a century or two of difficult empirical work. Still, there is every reason to believe that the methods of cognitive science and neuroscience will succeed.

The really hard problem of consciousness is the problem of *experience*. When we think and perceive, there is a whir of information-processing, but there is also a subjective aspect. As Nagel (1974) has put it, there is *something it is like* to be a conscious organism. This subjective aspect is experience. When we see, for example, we *experience* visual sensations: the felt quality of redness, the experience of dark and light, the quality of depth in a visual field. Other experiences go along with perception in different modalities: the sound of a clarinet, the smell of mothballs. Then there are bodily sensations, from pains to orgasms; mental images that are conjured up internally; the felt quality of emotion, and the experience of a stream of conscious thought. What unites all of these states is that there is something it is like to be in them. All of them are states of experience.

It is undeniable that some organisms are subjects of experience. But the question of how it is that these systems are subjects of experience is perplexing. Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does.

If any problem qualifies as *the* problem of consciousness, it is this one. In this central sense of "consciousness", an organism is conscious if there is something it is like to be that organism, and a mental state is conscious if there is something it is like to be in that state. Sometimes terms such as "phenomenal consciousness" and "qualia" are also used here, but I find it more natural to speak of "conscious experience" or simply "experience". Another useful way to avoid confusion (used by e.g. Newell 1990, Chalmers 1995) is to reserve the term "consciousness" for the phenomena of experience, using the less loaded term "awareness" for the more straightforward phenomena described earlier. If such a convention were widely adopted, communication would be much easier; as things stand, those who talk about "consciousness" are frequently talking past each other.

The ambiguity of the term "consciousness" is often exploited by both philosophers and scientists writing on the subject. It is common to see a paper on consciousness begin with an invocation of the mystery of consciousness, noting the strange intangibility and ineffability of subjectivity, and worrying that so far we have no theory of the phenomenon. Here, the topic is clearly the hard problem—the problem of experience. In the second half of the paper, the tone becomes more optimistic, and the author's own theory of consciousness is outlined. Upon examination, this theory turns out to be a theory of one of the more straightforward phenomena—of reportability, of introspective access, or whatever. At the close, the author declares that consciousness has turned out to be tractable after all, but the reader is left feeling like the victim of a bait-and-switch. The hard problem remains untouched.

3 Functional explanation

Why are the easy problems easy, and why is the hard problem hard? The easy problems are easy precisely because they concern the explanation of cognitive *abilities* and *functions*. To explain a cognitive function, we need only specify a mechanism that can perform the function. The methods of cognitive science are well-suited for this sort of explanation, and so are well-suited to the easy problems of consciousness. By contrast, the hard problem is hard precisely because it is not a problem about the performance of functions. The problem persists even when the performance of all the relevant functions is explained. (Here "function" is not used in the narrow teleological sense of something that a system is designed to do, but in the broader sense of any causal role in the production of behavior that a system might perform.)

To explain reportability, for instance, is just to explain how a system could perform the function of producing reports on internal states. To explain internal access, we need to explain how a system could be appropriately affected by its internal states and use information about those states in directing later processes. To explain integration and control, we need to explain how a system's central processes can bring information contents together and use them in the facilitation of various behaviors. These are all problems about the explanation of functions.

How do we explain the performance of a function? By specifying a *mechanism* that performs the function. Here, neurophysiological and cognitive modeling are perfect for the task. If we want a detailed low-level explanation, we can specify the neural mechanism that is responsible for the function. If we want a more abstract explanation, we can specify a mechanism in computational terms. Either way, a full and satisfying explanation will result. Once we have specified the neural or computational mechanism that performs the function of verbal report, for example, the bulk of our work in explaining reportability is over.

In a way, the point is trivial. It is a *conceptual* fact about these phenomena that their explanation only involves the explanation of various functions, as the phenomena are *functionally definable*. All it *means* for reportability to be instantiated in a system is that the system has the capacity for verbal reports of internal information. All it means for a system to be awake is for it to be appropriately receptive to information from the environment and for it to be able to use this information in directing behavior in an appropriate way. To see that this sort of thing is a conceptual fact, note that someone who says "you have explained the performance of the verbal report function, but you have not explained reportability" is making a trivial conceptual mistake about reportability. All it could *possibly* take to explain reportability is an explanation of how the relevant function is performed; the same goes for the other phenomena in question.

Throughout the higher-level sciences, reductive explanation works in just this way. To explain the gene, for instance, we needed to specify the mechanism that stores and transmits hereditary information from one generation to the next. It turns out that DNA performs this function; once we explain how the function is performed, we have explained the gene. To explain life, we ultimately need to explain how a system can reproduce, adapt to its environment, metabolize, and so on. All of these are questions about the performance of functions, and so are well-suited to reductive explanation. The same holds for most problems in cognitive science. To explain learning, we need to explain the way in which a system's behavioral capacities are modified in light of environmental information, and the way in which new information can be brought to bear in adapting a system's actions to its environment. If we show how a neural or computational mechanism does the job, we have explained learning. We can say the same for other cognitive phenomena, such as perception, memory, and language. Sometimes the relevant functions need to be characterized quite subtly, but it is clear that insofar as cognitive science explains these phenomena at all, it does so by explaining the performance of functions.

When it comes to conscious experience, this sort of explanation fails. What makes the hard problem hard and almost unique is that it goes *beyond* problems about the performance of functions. To see this, note that even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience—perceptual discrimination, categorization, internal access, verbal report—there may still remain a further unanswered question: *Why is the performance of these functions accompanied by experience?* A simple explanation of the functions leaves this question open.

There is no analogous further question in the explanation of genes, or of life, or of learning. If someone says "I can see that you have explained how DNA stores and transmits hereditary information from one generation to the next, but you have not explained how it is a *gene*", then they are making a conceptual mistake. All it means to be a gene is to be an entity that performs the relevant storage and transmission function. But if someone says "I can see that you have explained how information is discriminated, integrated, and reported, but you have not explained how it is *experienced*", they are not making a conceptual mistake. This is a nontrivial further question.

This further question is the key question in the problem of consciousness. Why doesn't all this information-processing go on "in the dark", free of any inner feel? Why is it that when electromagnetic waveforms impinge on a retina and are discriminated and categorized by a visual system, this discrimination and categorization is experienced as a sensation of vivid red? We know that conscious experience *does* arise when these functions are performed, but the very fact that it arises is the central mystery. There is an *explanatory gap* (a term due to Levine 1983) between the functions and experience, and we need an explanatory bridge to cross it. A mere account of the functions stays on one side of the gap, so the materials for the bridge must be found elsewhere.

This is not to say that experience *has* no function. Perhaps it will turn out to play an important cognitive role. But for any role it might play, there will be more to the explanation of experience than a simple explanation of the function. Perhaps it will even turn out that in the course of explaining a function, we will be led to the key insight that allows an explanation of experience. If this happens, though, the discovery will be an *extra* explanatory reward. There is no cognitive function such that we can say in advance that explanation of that function will *automatically* explain experience.

To explain experience, we need a new approach. The usual explanatory methods of cognitive science and neuroscience do not suffice. These methods have been developed precisely to explain the performance of cognitive functions, and they do a good job of it. But as these methods stand, they are *only* equipped to explain the performance of functions. When it comes to the hard problem, the standard approach has nothing to say.

4 Some case-studies

In the last few years, a number of works have addressed the problems of consciousness within the framework of cognitive science and neuroscience. This might suggest that the analysis above is faulty, but in fact a close examination of the relevant work only lends the analysis further support.

When we investigate just which aspects of consciousness these studies are aimed at, and which aspects they end up explaining, we find that the ultimate target of explanation is always one of the easy problems. I will illustrate this with two representative examples.

The first is the "neurobiological theory of consciousness" outlined by Crick and Koch (1990; see also Crick 1994). This theory centers on certain 35-75 hertz neural oscillations in the cerebral cortex; Crick and Koch hypothesize that these oscillations are the basis of consciousness. This is partly because the oscillations seem to be correlated with awareness in a number of different modalities—within the visual and olfactory systems, for example—and also because they suggest a mechanism by which the *binding* of information contents might be achieved. Binding is the process whereby separately represented pieces of information about a single entity are brought together to be used by later processing, as when information about the color and shape of a perceived object is integrated from separate visual pathways. Following others (e.g., Eckhorn *et al* 1988), Crick and Koch hypothesize that binding may be achieved by the synchronized oscillations of neuronal groups representing the relevant contents. When two pieces of information are to be bound together, the relevant neural groups will oscillate with the same frequency and phase.

The details of how this binding might be achieved are still poorly understood, but suppose that they can be worked out. What might the resulting theory explain? Clearly it might explain the binding of information contents, and perhaps it might yield a more general account of the integration of information in the brain. Crick and Koch also suggest that these oscillations activate the mechanisms of working memory, so that there may be an account of this and perhaps other forms of memory in the distance. The theory might eventually lead to a general account of how perceived information is bound and stored in memory, for use by later processing.

Such a theory would be valuable, but it would tell us nothing about why the relevant contents are experienced. Crick and Koch suggest that these oscillations are the neural *correlates* of experience. This claim is arguable—does not binding also take place in the processing of unconscious information?—but even if it is accepted, the *explanatory* question remains: Why do the oscillations give rise to experience? The only basis for an explanatory connection is the role they play in binding and storage, but the question of why binding and storage should themselves be accompanied by experience is never addressed. If we do not know why binding and storage should give rise to experience, telling a story about the oscillations cannot help us. Conversely, if we *knew* why binding and storage gave rise to experience, the neurophysiological details would be just the icing on the cake. Crick and Koch's theory gains its purchase by *assuming* a connection between binding and experience, and so can do nothing to explain that link.

I do not think that Crick and Koch are ultimately claiming to address the hard problem, although some have interpreted them otherwise. A published interview with Koch gives a clear statement of the limitations on the theory's ambitions.

> Well, let's first forget about the really difficult aspects, like subjective feelings, for they may not have a scientific solution. The subjective state of play, of pain, of plea-

> sure, 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." ("What is Consciousness", *Discover*, November 1992, p. 96.)

The second example is an approach at the level of cognitive psychology. This is Baars' global workspace theory of consciousness, presented in his book *A Cognitive Theory of Consciousness*. According to this theory, the contents of consciousness are contained in a *global workspace*, a central processor used to mediate communication between a host of specialized nonconscious processors. When these specialized processors need to broadcast information to the rest of the system, they do so by sending this information to the workspace, which acts as a kind of communal blackboard for the rest of the system, accessible to all the other processors.

Baars uses this model to address many aspects of human cognition, and to explain a number of contrasts between conscious and unconscious cognitive functioning. Ultimately, however, it is a theory of *cognitive accessibility*, explaining how it is that certain information contents are widely accessible within a system, as well as a theory of informational integration and reportability. The theory shows promise as a theory of awareness, the functional correlate of conscious experience, but an explanation of experience itself is not on offer.

One might suppose that according to this theory, the contents of experience are precisely the contents of the workspace. But even if this is so, nothing internal to the theory *explains* why the information within the global workspace is experienced. The best the theory can do is to say that the information is experienced because it is *globally accessible*. But now the question arises in a different form: why should global accessibility give rise to conscious experience? As always, this bridging question is unanswered.

Almost all work taking a cognitive or neuroscientific approach to consciousness in recent years could be subjected to a similar critique. The "Neural Darwinism" model of Edelman (1989), for instance, addresses questions about perceptual awareness and the self-concept, but says nothing about why there should also be experience. The "multiple drafts" model of Dennett (1991) is largely directed at explaining the reportability of certain mental contents. The "intermediate level" theory of Jackendoff (1988) provides an account of some computational processes that underlie consciousness, but Jackendoff stresses that the question of how these "project" into conscious experience remains mysterious.

Researchers using these methods are often inexplicit about their attitudes to the problem of conscious experience, although sometimes they take a clear stand. Even among those who are clear about it, attitudes differ widely. In placing this sort of work with respect to the problem of experience, a number of different strategies are available. It would be useful if these strategic choices were more often made explicit.

The first strategy is simply to *explain something else*. Some researchers are explicit that the problem of experience is too difficult for now, and perhaps even outside the domain of science altogether. These researchers instead choose to address one of the more tractable problems such as reportability or the self-concept. Although I have called these problems the "easy" problems, they are among the most interesting unsolved problems in cognitive science, so this work is certainly worthwhile. The worst that can be said of this choice is that in the context of research on consciousness it is relatively unambitious, and the work can sometimes be misinterpreted.

The second choice is to take a harder line and *deny the phenomenon*. (Variations on this approach are taken by Allport 1988, Dennett 1991, and Wilkes 1988.) According to this line, once we have explained the functions such as accessibility, reportability, and the like, there is no further phenomenon called "experience" to explain. Some explicitly deny the phenomenon, holding for example that what is not externally verifiable cannot be real. Others achieve the same effect by allowing that experience exists, but only if we equate "experience" with something like the capacity to discriminate and report. These approaches lead to a simpler theory, but are ultimately unsatisfactory. Experience is the most central and manifest aspect of our mental lives, and indeed is perhaps the key explanandum in the science of the mind. Because of this status as an explanandum, experience cannot be discarded like the vital spirit when a new theory comes along. Rather, it is the central fact that any theory of consciousness must explain. A theory that denies the phenomenon "solves" the problem by ducking the question.

In a third option, some researchers *claim to be explaining experience* in the full sense. These researchers (unlike those above) wish to take experience very seriously; they lay out their functional model or theory, and claim that it explains the full subjective quality of experience (e.g. Flohr 1992, Humphrey 1992). The relevant step in the explanation is usually passed over quickly, however, and usually ends up looking something like magic. After some details about information processing are given, experience suddenly enters the picture, but it is left obscure *how* these processes should suddenly give rise to experience. Perhaps it is simply taken for granted that it does, but then we have an incomplete explanation and a version of the fifth strategy below.

A fourth, more promising approach appeals to these methods to *explain the structure of experience*. For example, it is arguable that an account of the discriminations made by the visual system can account for the structural relations between different color experiences, as well as for the geometric structure of the visual field (see e.g., Clark 1992 and Hardin 1992). In general, certain facts about structures found in processing will correspond to and arguably explain facts about the structure of experience. This strategy is plausible but limited. At best, it takes the existence of experience for granted and accounts for some facts about its structure, providing a sort of nonreductive explanation of the structural aspects of experience (I will say more on this later). This is useful for many purposes, but it tells us nothing about why there should be experience in the first place.

A fifth and reasonable strategy is to *isolate the substrate of experience*. After all, almost everyone allows that experience *arises* one way or another from brain processes, and it makes sense to identify the sort of process from which it arises. Crick and Koch put their work forward as isolating the neural correlate of consciousness, for example, and Edelman (1989) and Jackendoff (1988) make similar claims. Justification of these claims requires a careful theoretical analysis, especially as experience is not directly observable in experimental contexts, but when applied judiciously this strategy can shed indirect light on the problem of experience. Nevertheless, the strategy is clearly incomplete. For a satisfactory theory, we need to know more than *which* processes give rise to experience; we need an account of why and how. A full theory of consciousness must build an explanatory bridge.

5 The extra ingredient

We have seen that there are systematic reasons why the usual methods of cognitive science and neuroscience fail to account for conscious experience. These are simply the wrong sort of methods: nothing that they give to us can yield an explanation. To account for conscious experience, we need an *extra ingredient* in the explanation. This makes for a *challenge* to those who are serious about the hard problem of consciousness: What is your extra ingredient, and why should *that* account for conscious experience?

There is no shortage of extra ingredients to be had. Some propose an injection of chaos and nonlinear dynamics. Some think that the key lies in nonalgorithmic processing. Some appeal to future discoveries in neurophysiology. Some suppose that the key to the mystery will lie at the level of quantum mechanics. It is easy to see why all these suggestions are put forward. None of the old methods work, so the solution must lie with *something* new. Unfortunately, these suggestions all suffer from the same old problems.

Nonalgorithmic processing, for example, is put forward by Penrose (1989; 1994) because of the role it might play in the process of conscious mathematical insight. The arguments about mathematics are controversial, but even if they succeed and an account of nonalgorithmic processing in the human brain is given, it will still only be an account of the *functions* involved in mathematical reasoning and the like. For a nonalgorithmic process as much as an algorithmic process, the question is left unanswered: why should this process give rise to experience? In answering *this* question, there is no special role for nonalgorithmic processing.

The same goes for nonlinear and chaotic dynamics. These might provide a novel account of the dynamics of cognitive functioning, quite different from that given by standard methods in cognitive science. But from dynamics, one only gets more dynamics. The question about experience here is as mysterious as ever. The point is even clearer for new discoveries in neurophysiology. These new discoveries may help us make significant progress in understanding brain function, but for any neural process we isolate, the same question will always arise. It is difficult to imagine what a proponent of new neurophysiology expects to happen, over and above the explanation of further cognitive functions. It is not as if we will suddenly discover a phenomenal glow inside a neuron!

Perhaps the most popular "extra ingredient" of all is quantum mechanics (e.g. Hameroff 1994). The attractiveness of quantum theories of consciousness may stem from a Law of Minimization of Mystery: consciousness is mysterious and quantum mechanics is mysterious, so maybe the two mysteries have a common source. Nevertheless, quantum theories of consciousness suffer from the same difficulties as neural or computational theories. Quantum phenomena have some remarkable functional properties, such as nondeterminism and nonlocality. It is natural to speculate that these properties may play some role in the explanation of cognitive functions, such as random choice and the integration of information, and this hypothesis cannot be ruled out *a priori*. But when it comes to the explanation of experience, quantum processes are in the same boat as any other. The question of why these processes should give rise to experience is entirely unanswered.

(One special attraction of quantum theories is the fact that on some interpretations of quantum mechanics, consciousness plays an active role in "collapsing" the quantum wave function. Such interpretations are controversial, but in any case they offer no hope of *explaining* consciousness in terms of quantum processes. Rather, these theories *assume* the existence of consciousness, and use it in the explanation of quantum processes. At best, these theories tell us something about a physical role that consciousness may play. They tell us nothing about how it arises.)

At the end of the day, the same criticism applies to *any* purely physical account of consciousness. For any physical process we specify there will be an unanswered question: Why should this process give rise to experience? Given any such process, it is conceptually coherent that it could be instantiated in the absence of experience. It follows that no mere account of the physical process will tell us why experience arises. The emergence of experience goes beyond what can be derived from physical theory.

Purely physical explanation is well-suited to the explanation of physical *structures*, explaining macroscopic structures in terms of detailed microstructural constituents; and it provides a satisfying explanation of the performance of *functions*, accounting for these functions in terms of the physical mechanisms that perform them. This is because a physical account can *entail* the facts about structures and functions: once the internal details of the physical account are given, the structural and functional properties fall out as an automatic consequence. But the structure and dynamics of physical processes yield only more structure and dynamics, so structures and functions are all we can expect these processes to explain. The facts about experience cannot be an automatic consequence of any physical account, as it is conceptually coherent that any given process could exist without experience. Experience may *arise* from the physical, but it is not *entailed* by the physical.

The moral of all this is that *you can't explain conscious experience on the cheap*. It is a remarkable fact that reductive methods—methods that explain a high-level phenomenon wholly in terms of more basic physical processes—work well in so many domains. In a sense, one *can* explain most biological and cognitive phenomena on the cheap, in that these phenomena are seen as automatic consequences of more fundamental processes. It would be wonderful if reductive methods could explain experience, too (I hoped for a long time that they might). Unfortunately, there are systematic reasons why these methods must fail. Reductive methods are successful in most domains because what needs explaining in those domains are structures and functions, and these are the kind of thing that a physical account can entail. When it comes to a problem over and above the explanation of structures and functions, these methods are impotent.

This might seem reminiscent of the vitalist claim that no physical account could explain life, but the cases are disanalogous. What drove vitalist skepticism was doubt about whether physical mechanisms could perform the many remarkable functions associated with life, such as complex adaptive behavior and reproduction. The conceptual claim that explanation of functions is what is needed was implicitly accepted, but lacking detailed knowledge of biochemical mechanisms, vitalists doubted whether any physical process could do the job and put forward the hypothesis of the vital spirit as an alternative explanation. Once it turned out that physical processes could perform the relevant functions, vitalist doubts melted away.

With experience, on the other hand, physical explanation of the functions is not in question. The key is instead the *conceptual* point that the explanation of functions does not suffice for the explanation of experience. This basic conceptual point is not something that further neuroscientific investigation will affect. In a similar way, experience is disanalogous to the *élan vital*. The vital spirit was put forward as an explanatory posit, in order to explain the relevant functions, and could therefore be discarded when those functions were explained without it. Experience is not an explanatory posit but an explanandum in its own right, and so is not a candidate for this sort of elimination.

It is tempting to note that all sorts of puzzling phenomena have eventually turned out to be explainable in physical terms. But each of these were problems about the observable behavior of physical objects, coming down to problems in the explanation of structures and functions. Because of this, these phenomena have always been the kind of thing that a physical account *might* explain, even if at some points there have been good reasons to suspect that no such explanation would be forthcoming. The tempting induction from these cases fails in the case of consciousness, which is not a problem about physical structures and functions. The problem of consciousness is puzzling in an entirely different way. An analysis of the problem shows us that conscious experience is just not the kind of thing that a wholly reductive account could succeed in explaining.

6 Nonreductive explanation

At this point some are tempted to give up, holding that we will never have a theory of conscious experience. McGinn (1989), for example, argues that the problem is too hard for our limited minds; we are "cognitively closed" with respect to the phenomenon. Others have argued that conscious experience lies outside the domain of scientific theory altogether.

I think this pessimism is premature. This is not the place to give up; it is the place where things get interesting. When simple methods of explanation are ruled out, we need to investigate the alternatives. Given that reductive explanation fails, *nonreductive* explanation is the natural choice.

Although a remarkable number of phenomena have turned out to be explicable wholly in terms of entities simpler than themselves, this is not universal. In physics, it occasionally happens that an entity has to be taken as *fundamental*. Fundamental entities are not explained in terms of anything simpler. Instead, one takes them as basic, and gives a theory of how they relate to everything else in the world. For example, in the nineteenth century it turned out that electromagnetic processes could not be explained in terms of the wholly mechanical processes that previous physical theories appealed to, so Maxwell and others introduced electromagnetic charge and electromagnetic forces as new fundamental components of a physical theory. To explain electromagnetism, the ontology of physics had to be expanded. New basic properties and basic laws were needed to give a satisfactory account of the phenomena.

Other features that physical theory takes as fundamental include mass and space-time. No attempt is made to explain these features in terms of anything simpler. But this does not rule out the possibility of a theory of mass or of space-time. There is an intricate theory of how these features interrelate, and of the basic laws they enter into. These basic principles are used to explain many familiar phenomena concerning mass, space, and time at a higher level.

I suggest that a theory of consciousness should take experience as fundamental. We know that a theory of consciousness requires the addition of *something* fundamental to our ontology, as everything in physical theory is compatible with the absence of consciousness. We might add some entirely new nonphysical feature, from which experience can be derived, but it is hard to see what such a feature would be like. More likely, we will take experience itself as a fundamental feature of the world, alongside mass, charge, and space-time. If we take experience as fundamental, then we can go about the business of constructing a theory of experience.

Where there is a fundamental property, there are fundamental laws. A nonreductive theory of experience will add new principles to the furniture of the basic laws of nature. These basic principles will ultimately carry the explanatory burden in a theory of consciousness. Just as we explain familiar high-level phenomena involving mass in terms of more basic principles involving mass and other entities, we might explain familiar phenomena involving experience in terms of more basic principles involving experience and other entities.

In particular, a nonreductive theory of experience will specify basic principles telling us how experience depends on physical features of the world. These *psychophysical* principles will not interfere with physical laws, as it seems that physical laws already form a closed system. Rather, they will be a supplement to a physical theory. A physical theory gives a theory of physical processes, and a psychophysical theory tells us how those processes give rise to experience. We know that experience depends on physical processes, but we also know that this dependence cannot be derived from physical laws alone. The new basic principles postulated by a nonreductive theory give us the extra ingredient that we need to build an explanatory bridge.

Of course, by taking experience as fundamental, there is a sense in which this approach does not tell us why there is experience in the first place. But this is the same for any fundamental theory. Nothing in physics tells us why there is matter in the first place, but we do not count this against theories of matter. Certain features of the world need to be taken as fundamental by any scientific theory. A theory of matter can still explain all sorts of facts about matter, by showing how they are consequences of the basic laws. The same goes for a theory of experience.

This position qualifies as a variety of dualism, as it postulates basic properties over and above the properties invoked by physics. But it is an innocent version of dualism, entirely compatible with the scientific view of the world. Nothing in this approach contradicts anything in physical theory; we simply need to add further *bridging* principles to explain how experience arises from physical processes. There is nothing particularly spiritual or mystical about this theory—its overall shape is like that of a physical theory, with a few fundamental entities connected by fundamental laws. It expands the ontology slightly, to be sure, but Maxwell did the same thing. Indeed, the overall structure of this position is entirely naturalistic, allowing that ultimately the universe comes down to a network of basic entities obeying simple laws, and allowing that there may ultimately be a theory of consciousness cast in terms of such laws. If the position is to have a name, a good choice might be *naturalistic dualism*.

If this view is right, then in some ways a theory of consciousness will have more in common with a theory in physics than a theory in biology. Biological theories involve no principles that are fundamental in this way, so biological theory has a certain complexity and messiness to it; but theories in physics, insofar as they deal with fundamental principles, aspire to simplicity and elegance. The fundamental laws of nature are part of the basic furniture of the world, and physical theories are telling us that this basic furniture is remarkably simple. If a theory of consciousness also involves fundamental principles, then we should expect the same. The principles of simplicity, elegance, and even beauty that drive physicists' search for a fundamental theory will also apply to a theory of consciousness.

(A technical note: Some philosophers argue that even though there is a *conceptual* gap between physical processes and experience, there need be no metaphysical gap, so that experience might in a certain sense still be physical (e.g. Levine 1983; Loar 1990; Byrne 1993; Papineau 1994; Sturgeon 1994). Usually this line of argument is supported by an appeal to the notion of *a posteriori* necessity (Kripke 1980). I think that this position rests on a misunderstanding of *a posteriori* necessity, however, or else requires an entirely new sort of necessity that we have no reason to believe in; see Chalmers 1995 (also Jackson 1994 and Lewis 1994) for details. In any case, this position still concedes an *explanatory* gap between physical processes and experience. For example, the principles connecting the physical and the experiential will not be derivable from the laws of physics, so such principles must be taken as *explanatorily* fundamental. So even on this sort of view, the explanatory structure of a theory of consciousness will be much as I have described.)

7 Outline of a theory of consciousness

It is not too soon to begin work on a theory. We are already in a position to understand certain key facts about the relationship between physical processes and experience, and about the regularities that connect them. Once reductive explanation is set aside, we can lay those facts on the table so that they can play their proper role as the initial pieces in a nonreductive theory of consciousness, and as constraints on the basic laws that constitute an ultimate theory.

There is an obvious problem that plagues the development of a theory of consciousness, and that is the paucity of objective data. Conscious experience is not directly observable in an experimental context, so we cannot generate data about the relationship between physical processes and experience at will. Nevertheless, we all have access to a rich source of data in our own case. Many important regularities between experience and processing can be inferred from considerations about one's own experience. There are also good indirect sources of data from observable cases, as when one relies on the verbal report of a subject as an indication of experience. These methods have their limitations, but we have more than enough data to get a theory off the ground.

Philosophical analysis is also useful in getting value for money out of the data we have. This sort of analysis can yield a number of principles relating consciousness and cognition, thereby strongly constraining the shape of an ultimate theory. The method of thought-experimentation can also yield significant rewards, as we will see. Finally, the fact that we are searching for a *fundamental* theory means that we can appeal to such nonempirical constraints as simplicity, homogeneity, and the like in developing a theory. We must seek to systematize the information we have, to extend it as far as possible by careful analysis, and then make the inference to the simplest possible theory that explains the data while remaining a plausible candidate to be part of the fundamental furniture of the world.

Such theories will always retain an element of speculation that is not present in other scientific theories, because of the impossibility of conclusive intersubjective experimental tests. Still, we can certainly construct theories that are compatible with the data that we have, and evaluate them in comparison to each other. Even in the absence of intersubjective observation, there are numerous criteria available for the evaluation of such theories: simplicity, internal coherence, coherence with theories in other domains, the ability to reproduce the properties of experience that are familiar from our own case, and even an overall fit with the dictates of common sense. Perhaps there will be significant indeterminacies remaining even when all these constraints are applied, but we can at least develop plausible candidates. Only when candidate theories have been developed will we be able to evaluate them.

A nonreductive theory of consciousness will consist in a number of *psychophysical principles*, principles connecting the properties of physical processes to the properties of experience. We can think of these principles as encapsulating the way in which experience arises from the physical. Ultimately, these principles should tell us what sort of physical systems will have associated experiences, and for the systems that do, they should tell us what sort of physical properties are relevant to the emergence of experience, and just what sort of experience we should expect any given physical system to yield. This is a tall order, but there is no reason why we should not get started.

In what follows, I present my own candidates for the psychophysical principles that might go into a theory of consciousness. The first two of these are *nonbasic principles*—systematic connections between processing and experience at a relatively high level. These principles can play a significant role in developing and constraining a theory of consciousness, but they are not cast at a sufficiently fundamental level to qualify as truly basic laws. The final principle is my candidate for a *basic principle* that might form the cornerstone of a fundamental theory of consciousness. This final principle is particularly speculative, but it is the kind of speculation that is required if we are ever to have a satisfying theory of consciousness. I can present these principles only briefly here; I argue for them at much greater length in Chalmers (1995).

1. **The principle of structural coherence**. This is a principle of coherence between the *structure of consciousness* and the *structure of awareness*. Recall that "awareness" was used earlier to refer to the various functional phenomena that are associated with consciousness. I am now using it to refer to a somewhat more specific process in the cognitive underpinnings of experience. In particular, the contents of awareness are to be understood as those information contents that are accessible to central systems, and brought to bear in a widespread way in the control of behavior. Briefly put, we can think of awareness as *direct availability for global control*. To a first approximation, the contents of awareness are the contents that are directly accessible and potentially reportable, at least in a language-using system.

Awareness is a purely functional notion, but it is nevertheless intimately linked to conscious experience. In familiar cases, wherever we find consciousness, we find awareness. Wherever there is conscious experience, there is some corresponding information in the cognitive system that is available in the control of behavior, and available for verbal report. Conversely, it seems that whenever information is available for report and for global control, there is a corresponding conscious experience. Thus, there is a direct correspondence between consciousness and awareness.

The correspondence can be taken further. It is a central fact about experience that it has a complex structure. The visual field has a complex geometry, for instance. There are also relations of similarity and difference between experiences, and relations in such things as relative intensity. Every subject's experience can be at least partly characterized and decomposed in terms of these structural properties: similarity and difference relations, perceived location, relative intensity, geometric structure, and so on. It is also a central fact that to each of these structural features, there is a corresponding feature in the information-processing structure of awareness.

Take color sensations as an example. For every distinction between color experiences, there is a corresponding distinction in processing. The different phenomenal colors that we experience form a complex three-dimensional space, varying in hue, saturation, and intensity. The properties of this space can be recovered from information-processing considerations: examination of the visual systems shows that waveforms of light are discriminated and analyzed along three different axes, and it is this three-dimensional information that is relevant to later processing. The three-dimensional structure of phenomenal color space therefore corresponds directly to the three dimensional structure of visual awareness. This is precisely what we would expect. After all, every color distinction corresponds to some reportable information, and therefore to a distinction that is represented in the structure of processing.

In a more straightforward way, the geometric structure of the visual field is directly reflected in a structure that can be recovered from visual processing. Every geometric relation corresponds to something that can be reported and is therefore cognitively represented. If we were given only the story about information-processing in an agent's visual and cognitive system, we could not *directly* observe that agent's visual experiences, but we could nevertheless infer those experiences' structural properties.

In general, any information that is consciously experienced will also be cognitively represented. The fine-grained structure of the visual field will correspond to some fine-grained structure in visual processing. The same goes for experiences in other modalities, and even for nonsensory experiences. Internal mental images have geometric properties that are represented in processing. Even emotions have structural properties, such as relative intensity, that correspond directly to a structural property of processing; where there is greater intensity, we find a greater effect on later processes. In general, precisely because the structural properties of experience are accessible and reportable, those properties will be directly represented in the structure of awareness.

It is this isomorphism between the structures of consciousness and awareness that constitutes the principle of structural coherence. This principle reflects the central fact that even though cognitive processes do not conceptually entail facts about conscious experience, consciousness and cognition do not float free of one another but cohere in an intimate way.

This principle has its limits. It allows us to recover structural properties of experience from information-processing properties, but not all properties of experience are structural properties. There are properties of experience, such as the intrinsic nature of a sensation of red, that cannot be fully captured in a structural description. The very intelligibility of inverted spectrum scenarios, where experiences of red and green are inverted but all structural properties remain the same, show that structural properties constrain experience without exhausting it. Nevertheless, the very fact that we feel compelled to leave structural properties unaltered when we imagine experiences inverted between functionally identical systems shows how central the principle of structural coherence is to our conception of our mental lives. It is not a *logically* necessary principle, as after all we can imagine all the information processing occurring without any experience at all, but it is nevertheless a strong and familiar constraint on the psychophysical connection.

The principle of structural coherence allows for a very useful kind of indirect explanation of experience in terms of physical processes. For example, we can use facts about neural processing of visual information to indirectly explain the structure of color space. The facts about neural processing can entail and explain the structure of awareness; if we take the coherence principle for granted, the structure of experience will also be explained. Empirical investigation might even lead us to better understand the structure of awareness within a bat, shedding indirect light on Nagel's vexing question of what it is like to be a bat. This principle provides a natural interpretation of much existing work on the explanation of consciousness (e.g. Clark 1992 and Hardin 1992 on colors, and Akins 1993 on bats), although it is often appealed to inexplicitly. It is so familiar that it is taken for granted by almost everybody, and is a central plank in the cognitive explanation of consciousness.

The coherence between consciousness and awareness also allows a natural interpretation of work in neuroscience directed at isolating the *substrate* (or the *neural correlate*) of consciousness. Various specific hypotheses have been put forward. For example, Crick and Koch (1990) suggest that 40-Hz oscillations may be the neural correlate of consciousness, whereas Libet (1993) suggests that temporally-extended neural activity is central. If we accept the principle of coherence, the most *direct* physical correlate of consciousness is awareness: the process whereby information is made directly available for global control. The different specific hypotheses can be interpreted as empirical suggestions about how awareness might be achieved. For example, Crick and Koch suggest that 40-Hz oscillations are the gateway by which information is integrated into working memory and thereby made available to later processes. Similarly, it is natural to suppose that Libet's temporally extended activity is relevant precisely because only that sort of activity achieves global availability. The same applies to other suggested correlates such as the "global workspace" of Baars (1988), the "high-quality representations" of Farah (1994), and the "selector inputs to action systems" of Shallice (1972). All these can be seen as hypotheses about the *mechanisms of awareness*: the mechanisms that perform the function of making information directly available for global control.

Given the coherence between consciousness and awareness, it follows that a mechanism of awareness will itself be a correlate of conscious experience. The question of just *which* mechanisms in the brain govern global availability is an empirical one; perhaps there are many such mechanisms. But if we accept the coherence principle, we have reason to believe that the processes that *explain* awareness will at the same time be part of the *basis* of consciousness.

2. **The principle of organizational invariance**. This principle states that any two systems with the same fine-grained *functional organization* will have qualitatively identical experiences. If the causal patterns of neural organization were duplicated in silicon, for example, with a silicon chip for every neuron and the same patterns of interaction, then the same experiences would arise. According to this principle, what matters for the emergence of experience is not the specific physical makeup of a system, but the abstract pattern of causal interaction between its components. This principle is controversial, of course. Some (e.g. Searle 1980) have thought that consciousness is tied to a specific biology, so that a silicon isomorph of a human need not be conscious. I believe that the principle can be given significant support by the analysis of thought-experiments, however.

Very briefly: suppose (for the purposes of a *reductio ad absurdum*) that the principle is false, and that there could be two functionally isomorphic systems with different experiences. Perhaps only one of the systems is conscious, or perhaps both are conscious but they have different experiences. For the purposes of illustration, let us say that one system is made of neurons and the other of silicon, and that one experiences red where the other experiences blue. The two systems have the same organization, so we can imagine gradually transforming one into the other, perhaps replacing neurons one at a time by silicon chips with the same local function. We thus gain a spectrum of intermediate cases, each with the same organization, but with slightly different physical makeup and slightly different experiences. Along this spectrum, there must be two systems A and B between which we replace less than one tenth of the system, but whose experiences differ. These two systems are physically identical, except that a small neural circuit in A has been replaced by a silicon circuit in B.

The key step in the thought-experiment is to take the relevant neural circuit in A, and install alongside it a causally isomorphic silicon circuit, with a switch between the two. What happens when we flip the switch? By hypothesis, the system's conscious experiences will change; from red to blue, say, for the purposes of illustration. This follows from the fact that the system after the change is essentially a version of B, whereas before the change it is just A.

But given the assumptions, there is no way for the system to *notice* the changes! Its causal organization stays constant, so that all of its functional states and behavioral dispositions stay fixed. As far as the system is concerned, nothing unusual has happened. There is no room for the thought, "Hmm! Something strange just happened!". In general, the structure of any such thought must be reflected in processing, but the structure of processing remains constant here. If there were to be such a thought it must float entirely free of the system and would be utterly impotent to affect later processing. (If it affected later processing, the systems would be functionally distinct, contrary to hypothesis). We might even flip the switch a number of times, so that experiences of red and blue dance back and forth before the system's "inner eye". According to hypothesis, the system can never notice these "dancing qualia".

This I take to be a *reductio* of the original assumption. It is a central fact about experience, very familiar from our own case, that whenever experiences change significantly and we are paying attention, we can notice the change; if this were not to be the case, we would be led to the skeptical possibility that our experiences are dancing before our eyes all the time. This hypothesis has the same status as the possibility that the world was created five minutes ago: perhaps it is logically coherent, but it is not plausible. Given the extremely plausible assumption that changes in experience correspond to changes in processing, we are led to the conclusion that the original hypothesis is impossible, and that any two functionally isomorphic systems must have the same sort of experiences. To put it in technical terms, the philosophical hypotheses of "absent qualia" and "inverted qualia", while logically possible, are empirically and nomologically impossible.

(Some may worry that a silicon isomorph of a neural system might be impossible for technical reasons. That question is open. The invariance principle says only that *if* an isomorph is possible, then it will have the same sort of conscious experience.)

There is more to be said here, but this gives the basic flavor. Once again, this thought experiment draws on familiar facts about the coherence between consciousness and cognitive processing to yield a strong conclusion about the relation between physical structure and experience. If the argument goes through, we know that the only physical properties directly relevant to the emergence of experience are *organizational* properties. This acts as a further strong constraint on a theory of consciousness.

3. **The double-aspect theory of information**. The two preceding principles have been *non-basic* principles. They involve high-level notions such as "awareness" and "organization", and therefore lie at the wrong level to constitute the fundamental laws in a theory of consciousness. Nevertheless, they act as strong constraints. What is further needed are *basic* principles that fit these constraints and that might ultimately explain them.

The basic principle that I suggest centrally involves the notion of *information*. I understand information in more or less the sense of Shannon (1948). Where there is information, there are *information states* embedded in an *information space*. An information space has a basic structure of *difference* relations between its elements, characterizing the ways in which different elements in a space are similar or different, possibly in complex ways. An information space is an abstract object, but following Shannon we can see information as *physically embodied* when there is a space of distinct physical states, the differences between which can be transmitted down some causal pathway. The states that are transmitted can be seen as themselves constituting an information space. To borrow a phrase from Bateson (1972), physical information is a *difference that makes a difference*.

The double-aspect principle stems from the observation that there is a direct isomorphism between certain physically embodied information spaces and certain *phenomenal* (or experiential) information spaces. From the same sort of observations that went into the principle of structural coherence, we can note that the differences between phenomenal states have a structure that corresponds directly to the differences embedded in physical processes; in particular, to those differences that make a difference down certain causal pathways implicated in global availability and control. That is, we can find the *same* abstract information space embedded in physical processing and in conscious experience.

This leads to a natural hypothesis: that information (or at least some information) has two basic aspects, a physical aspect and a phenomenal aspect. This has the status of a basic principle that might underlie and explain the emergence of experience from the physical. Experience arises by virtue of its status of one aspect of information, when the other aspect is found embodied in physical processing.

This principle is lent support by a number of considerations, which I can only outline briefly here. First, consideration of the sort of physical changes that correspond to changes in conscious experience suggests that such changes are always relevant by virtue of their role in constituting *informational changes*—differences within an abstract space of states that are divided up precisely according to their causal differences along certain causal pathways. Second, if the principle of organizational invariance is to hold, then we need to find some fundamental *organizational* property for experience to be linked to, and information is an organizational property *par excellence*. Third, this principle offers some hope of explaining the principle of structural coherence in terms of the structure present within information spaces. Fourth, analysis of the cognitive explanation of our *judgments* and *claims* about conscious experience—judgments that are functionally explainable but nevertheless deeply tied to experience itself—suggests that explanation centrally involves the information states embedded in cognitive processing. It follows that a theory based on information allows a deep coherence between the explanation of experience and the explanation of our judgments and claims about it.

Wheeler (1990) has suggested that information is fundamental to the physics of the universe. According to this "it from bit" doctrine, the laws of physics can be cast in terms of information, postulating different states that give rise to different effects without actually saying what those states *are*. It is only their position in an information space that counts. If so, then information is a natural candidate to also play a role in a fundamental theory of consciousness. We are led to a conception of the world on which information is truly fundamental, and on which it has two basic aspects, corresponding to the physical and the phenomenal features of the world.

Of course, the double-aspect principle is extremely speculative and is also underdetermined, leaving a number of key questions unanswered. An obvious question is whether *all* information has a phenomenal aspect. One possibility is that we need a further constraint on the fundamental theory, indicating just what *sort* of information has a phenomenal aspect. The other possibility is that there is no such constraint. If not, then experience is much more widespread than we might have believed, as information is everywhere. This is counterintuitive at first, but on reflection I think the position gains a certain plausibility and elegance. Where there is simple information processing, there is simple experience, and where there is complex information processing, there is complex experience. A mouse has a simpler information-processing structure than a human, and has correspondingly simpler experience; perhaps a thermostat, a maximally simple information processing structure, might have maximally simple experience? Indeed, if experience is truly a fundamental property, it would be surprising for it to arise only every now and then; most fundamental properties are more evenly spread. In any case, this is very much an open question, but I believe that the position is not as implausible as it is often thought to be.

Once a fundamental link between information and experience is on the table, the door is opened to some grander metaphysical speculation concerning the nature of the world. For example, it is often noted that physics characterizes its basic entities only *extrinsically*, in terms of their relations to other entities, which are themselves characterized extrinsically, and so on. The intrinsic nature of physical entities is left aside. Some argue that no such intrinsic properties exist, but then one is left with a world that is pure causal flux (a pure flow of information) with no properties for the causation to relate. If one allows that intrinsic properties exist, a natural speculation given the above is that the intrinsic properties of the physical—the properties that causation ultimately relates—are themselves phenomenal properties. We might say that phenomenal properties are the internal aspect of information. This could answer a concern about the causal relevance of experience—a natural worry, given a picture on which the physical domain is causally closed, and on which experience is supplementary to the physical. The informational view allows us to understand how experience might have a subtle kind of causal relevance in virtue of its status as the intrinsic nature of the physical. This metaphysical speculation is probably best ignored for the purposes of developing a scientific theory, but in addressing some philosophical issues it is quite suggestive.

8 Conclusion

The theory I have presented is speculative, but it is a candidate theory. I suspect that the principles of structural coherence and organizational invariance will be planks in any satisfactory theory of consciousness; the status of the double-aspect theory of information is less certain. Indeed, right now it is more of an idea than a theory. To have any hope of eventual explanatory success, it will have to be specified more fully and fleshed out into a more powerful form. Still, reflection on just what is plausible and implausible about it, on where it works and where it fails, can only lead to a better theory.

Most existing theories of consciousness either deny the phenomenon, explain something else, or elevate the problem to an eternal mystery. I hope to have shown that it is possible to make progress on the problem even while taking it seriously. To make further progress, we will need further investigation, more refined theories, and more careful analysis. The hard problem is a hard problem, but there is no reason to believe that it will remain permanently unsolved.[1]

References

Akins, K. 1993. What is it like to be boring and myopic? In (B. Dahlbom, ed.) *Dennett and his Critics*. Blackwell.

Allport, A. 1988. What concept of consciousness? In (A. Marcel and E. Bisiach, eds.) *Consciousness in Contemporary Science*. Oxford University Press.

Baars, B.J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Bateson, G. 1972. *Steps to an Ecology of Mind*. Chandler Publishing.

Byrne, A. 1993. *The Emergent Mind*. Ph.D. Dissertation, Princeton University.

Chalmers, D.J. 1995. *Toward a Theory of Consciousness*. MIT Press.

Clark, A. 1992. *Sensory Qualities*. Clarendon.

Crick, F. and Koch, C. 1990. Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2:263-275.

Crick, F. 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. Scribners.

Dennett, D.C. 1991. *Consciousness Explained*. Little-Brown.

[1]The arguments in this paper are presented in greater depth in my book *Toward a Theory of Consciousness*. Thanks to Francis Crick, Peggy DesAutels, Matthew Elton, Liane Gabora, Christof Koch, Paul Rhodes, Gregg Rosenberg, and Sharon Wahl for their comments.

Edelman, G. 1989. *The Remembered Present: A Biological Theory of Consciousness*. Basic Books.

Farah, M.J. 1994. Visual perception and visual awareness after brain damage: A tutorial overview. In (C. Umilta and M. Moscovitch, eds.) *Consciousness and Unconscious Information Processing: Attention and Performance 15*. MIT Press.

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

Hameroff, S.R. 1994. Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies* 1:91-118.

Hardin, C.L. 1992. Physiology, phenomenology, and Spinoza's true colors. In (A. Beckermann, H. Flohr, and J. Kim, eds.) *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter.

Humphrey, N. 1992. *A History of the Mind*. Simon and Schuster.

Jackendoff, R. 1987. *Consciousness and the Computational Mind*. MIT Press.

Jackson, F. 1994. Finding the mind in the natural world. In (R. Casati, B. Smith, and S. White, eds.) *Philosophy and the Cognitive Sciences*.

Kripke, S. 1980. *Naming and Necessity*. Harvard University Press.

Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-61.

Lewis, D. 1994. Reduction of mind. In (S. Guttenplan, ed.) *A Companion to the Philosophy of Mind*. Blackwell.

Libet, B. 1993. The neural time factor in conscious and unconscious events. In (G.R. Block and J. Marsh, eds.) *Experimental and Theoretical Studies of Consciousness* (Ciba Foundation Symposium 174). John Wiley and Sons.

Loar, B. 1990. Phenomenal states. *Philosophical Perspectives* 4:81-108.

McGinn, C. 1989. Can we solve the mind–body problem? *Mind* 98:349-66.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 4:435-50.

Newell, A. 1990. *Unified Theories of Cognition*. Harvard University Press.

Papineau, D. 1993. Physicalism, consciousness, and the antipathetic fallacy. *Australasian Journal of Philosophy* 71:169-83.

Penrose, R. 1989. *The Emperor's New Mind*. Oxford University Press.

Penrose, R. 1994. *Shadows of the Mind*. Oxford University Press.

Searle, J.R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3:417-57.

Shallice, T. 1972. Dual functions of consciousness. *Psychological Review* 79:383-93.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379-423.

Sturgeon, S. 1994. The epistemic basis of subjectivity. *Journal of Philosophy* 91:221-35.

Wheeler, J.A. 1990. Information, physics, quantum: The search for links. In (W. Zurek, ed.) *Complexity, Entropy, and the Physics of Information*. Addison-Wesley.

Wilkes, K.V. 1988. —, Yishi, Duh, Um and consciousness. In (A. Marcel and E. Bisiach, eds.) *Consciousness in Contemporary Science*. Oxford University Press.