Can Consciousness be Reductively Explained?

Selection from David Chalmers, *The Conscious Mind*, Chapter 3. (OUP, 1996)

1. Is consciousness logically supervenient on the physical?

Almost everything in the world pan be explained in physical terms; it is natural to hope that consciousness might be explained this way, too. In this chapter, however, I will argue that consciousness escapes the net of reductive explanation. No explanation given wholly in physical terms can ever account for the emergence of conscious experience. This may seem to be a negative conclusion, but it leads to some strong positive consequences that I will bring out in later chapters.

To make the case against reductive explanation, we need to show that consciousness is not logically supervenient on the physical. In principle, we need to show that it does not supervene *globally*—that is, that all the microphysical facts in the world do not entail the facts about consciousness. In practice, it is easier to run the argument *locally*, arguing that in an individual, microphysical facts do not entail the facts about consciousness. When it comes to consciousness, local and global supervenience plausibly stand and fall together, so it does not matter which way we run the argument: if consciousness supervenes at all, it almost certainly supervenes locally. If this is disputed, however, all the arguments can be run at the global level with straightforward alterations.

How can we argue that consciousness is not logically supervenient on the physical? There are various ways. We can think about what is conceivable, in order to argue directly for the logical possibility of a situation in which the physical facts are the same but the facts

about experience are different. We can appeal, to epistemology, arguing that the right sort of link between knowledge of physical facts and knowledge of consciousness is absent. And we can appeal directly to the concept of consciousness, arguing that there is no analysis of the concept that could ground an entailment from the physical to the phenomenal. In what follows I will give arguments using all three of these strategies. The first two are essentially arguments from conceivability, the second two are arguments from epistemology, and the fifth is an argument from analysis. There is some element of redundancy among the five arguments, but together they make a strong case.

One can also do things more directly, making the case against reductive explanation without explicitly appealing to logical supervenience. I have taken that route elsewhere, but here I will give the more detailed analysis to allow a fuller case. All the same, the case against reductive explanation and the critique of existing reductive accounts (in section 2 onward) should make sense even without this analysis. Some readers might like to proceed there directly, at least on a first reading.

(A technical note: The burden of this chapter is to argue, in effect, that there is no a priori entailment from physical facts to phenomenal facts. The sort of necessity that defines the relevant supervenience relation is the a priori version of logical necessity, where primary intensions are central. . . . [T]his is the relation that is relevant to issues about explanation; matters of a posteriori necessity can be set to one side. In the next chapter, issues of ontology rather than explanation are central, and I argue separately that there is no a posteriori necessary connection between physical facts and phenomenal facts.)

Argument 1: The Logical Possibility of Zombies

The most obvious way (although not the only way) to investigate the logical supervenience of consciousness is to consider 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 can consider 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.

So let us consider my zombie twin. This creature is molecule for molecule identical to me, and identical in all the low-level properties postulated by a completed physics, but he lacks conscious experience entirely. (Some might prefer to call a zombie "it," but I use the personal pronoun; I have grown quite fond of my zombie twin.) 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 pleasant taste experience through munching on a chocolate bar, and feeling a dull aching sensation in my right shoulder.

What is going on in my zombie twin? He is physically identical to me, and we may as well suppose that he is embedded in an identical environment. He 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. . . . He will be perceiving the trees outside, in the functional sense, and tasting the chocolate, 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. He will even be "conscious" in the functional senses described earlier—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.

This sort of zombie is quite unlike the zombies found in Hollywood movies, which tend to have significant functional impairments. The sort of consciousness that Hollywood zombies most obviously lack is a psychological version: typically, they have little capacity for introspection and lack a refined ability to voluntarily control behavior. They may or may not lack phenomenal consciousness; as Block (1995) 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 the *phenomenal zombies*, which are physically and functionally identical, but which lack experience. (Perhaps it is not surprising that phenomenal zombies have not been popular in Hollywood, as there would be obvious problems with their depiction.)

The idea of zombies as I have described them is a strange one. For a start, it is unlikely that zombies are naturally possible. In the real world, it is likely that any replica of me would be conscious. For this reason, it is most natural to imagine unconscious creatures as physically different from conscious ones—exhibiting impaired behavior, for example. But the question is not whether it is plausible that zombies could exist in our world, or even whether the idea of a zombie replica is a natural one; the question is whether the notion of a zombie is conceptually coherent. The mere intelligibility of the notion is enough to establish the conclusion.

Arguing for a logical possibility is not entirely straightforward. How, for example, 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 its obvious coherence. It seems quite clear that there is no hidden contradiction lurking in the description.

I confess 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 a brute intuition, 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 someone truly believes that a mile-high unicycle is logically impossible, she must give us some idea of where a contradiction lies, whether explicit or implicit. If she cannot point out something about the intensions of the concepts "mile-high" and "unicycle" that might lead to a contradiction, then her 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 a thousandfold. 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 further onto the other side.

For example, we can indirectly support the claim that zombies are logically possible by considering *non-standard 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 a 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 implausible that a set-up like this would give rise to 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; I suspect that in fact it would. . . . All that matters here is that the idea 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. We can make a similar point by considering my silicon isomorph, who is organized like me but who has silicon chips where I have neurons. Whether such an isomorph would *in fact* be conscious is controversial, but it seems to most people that those who deny this are expressing a coherent possibility. From these cases it follows that the existence of my conscious experience is not logically entailed by the facts about my functional organization.

But given that it is conceptually coherent that the group-mind set-up or my silicon isomorph could lack conscious experience, it

follows that my zombie twin is an equally coherent possibility. For it is clear that there is no more of a conceptual entailment from biochemistry to consciousness than there is from silicon or from a group of homunculi. If the silicon isomorph without conscious experience is conceivable, we need only substitute neurons for silicon in the conception while leaving functional organization constant, and we have my zombie twin. Nothing in this substitution could force experience into the conception; these implementational differences are simply not the sort of thing that could be conceptually relevant to experience. So consciousness fails to logically supervene on the physical.

The argument for zombies can be made without an appeal to these non-standard realizations, but these have a heuristic value in eliminating a source of conceptual confusion. To some people, intuitions about the logical possibility of an unconscious physical replica seem less than clear at first, perhaps because the familiar co-occurrence of biochemistry and consciousness can lead one to suppose a conceptual connection. Considerations of the less familiar cases remove these empirical correlations from the picture, and therefore make judgments of logical possibility more straightforward. But once it is accepted that these nonconscious functional replicas are logically possible, the corresponding conclusion concerning a physical replica cannot be avoided.

Some may think that conceivability arguments are unreliable. For example, sometimes it is objected that we cannot really imagine in detail the many billions of neurons in the human brain. Of course this is true; but we do not need to imagine each of the neurons to make the case. Mere complexity among neurons could not conceptually entail consciousness; if all that neural structure is to be relevant to consciousness, it must be relevant *in virtue* of some higher-level properties that it enables. So it is enough to imagine the system at a coarse level, and to make sure that we conceive it with appropriately sophisticated mechanisms of perception, categorization, high-bandwidth access to information contents, reportability, and the like. No matter how sophisticated we imagine

these mechanisms to be, the zombie scenario remains as coherent as ever. Perhaps an opponent might claim that all the unimagined neural detail is conceptually relevant in some way independent of its contribution to sophisticated functioning; but then she owes us an account of what that way might be, and none is available. Those implementational details simply lie at the wrong level to be conceptually relevant to consciousness.

It is also sometimes said that conceivability is an imperfect guide to possibility. The main way that conceivability and possibility can come apart is tied to the phenomenon of a posteriori necessity: for example, the hypothesis that water is not H_2O seems conceptually coherent, but water is arguably H_2O in all possible worlds. But a posteriori necessity is irrelevant to the concerns of this chapter. As we saw in the last chapter, explanatory connections are grounded in a priori entailments from physical facts to high-level facts. The relevant kind of possibility is to be evaluated using the primary intensions of the terms involved, instead of the secondary intensions that are relevant to a posteriori necessity. So even if a zombie world is conceivable only in the sense in which it is conceivable that water is not H_2O, that is enough to establish that consciousness cannot be reductively explained.

Those considerations aside, the main way in which conceivability arguments can go wrong is by subtle conceptual confusion: if we are insufficiently reflective we can overlook an incoherence in a purported possibility, by taking a conceived-of situation and *misdescribing* it. For example, one might think that one can conceive of a situation in which Fermat's last theorem is false, by imagining a situation in which leading mathematicians declare that they have found a counterexample. But given that the theorem is actually true, this situation is being misdescribed: it is really a scenario in which Fermat's last theorem is true, and in which some mathematicians make a mistake. Importantly, though, this kind of mistake always lies in the a priori domain, as it arises from the incorrect application of the primary intensions of our concepts to a conceived situation. Sufficient reflection will reveal that the

concepts are being incorrectly applied, and that the claim of logical possibility is not justified.

So the only route available to an opponent here is to claim that in describing the zombie world as a zombie world, we are misapplying the concepts, and that in fact there is a conceptual contradiction lurking in the description. Perhaps if we thought about it clearly enough we would realize that by imagining a physically identical world we are thereby *automatically* imagining a world in which there is conscious experience. But then the burden is on the opponent to give us some idea of where the contradiction might lie in the apparently quite coherent description. If no internal incoherence can be revealed, then there is a very strong case that the zombie world is logically possible.

As before, I can detect no internal incoherence; I have a clear picture of what I am conceiving when I conceive of a zombie. Still, some people find conceivability arguments difficult to adjudicate, particularly where strange ideas such as this one are concerned. It is therefore fortunate that every point made using zombies can also be made in other ways, for example by considering epistemology and analysis. To many, arguments of the latter sort (such as arguments 3-5 below) are more straightforward and therefore make a stronger foundation in the argument against logical supervenience. But zombies at least provide a vivid illustration of important issues in the vicinity. . . .

Argument 3: From Epistemic Asymmetry

As we saw earlier, consciousness is a surprising feature of the universe. 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, causation, and evolution among all the fields and particles in the spatiotemporal manifold—*that* information would not lead us to postulate the existence of conscious experience. My knowledge of

consciousness, in the first instance, comes from my own case, not from any external observation. It is my first-person 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 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 might be most natural to draw an eliminativist conclusion—that there was in fact no *experience* present in these creatures, just a lot of talk.

Eliminativism about conscious experience is an unreasonable position *only* because of our own acquaintance with 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 primarily from our own case, with external evidence playing at best a secondary role.

The point can also be made pointing to the existence of a problem of other minds. Even when we know everything physical about the other creatures we do not *know* for certain that they are conscious, or what their experiences are (although we may give good reason to believe that they are). It is striking that there is no problem of

"other lives," or of "other economies," or of "other heights." There is no epistemic asymmetry in those cases, precisely because those phenomena are logically supervenient on the physical.

The epistemic asymmetry in knowledge of consciousness makes it clear that consciousness cannot logically supervene. If it were logically supervenient, there would be no such epistemic asymmetry; 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—memory, perhaps—that are more easily detected in the first-person case. But this is just a matter of how hard one has to work. The presence of memory is just as accessible from the third person, in principle, as from the first person. The epistemic asymmetry associated with consciousness is much more fundamental, and it tells us that no collection of facts about complex causation in physical systems adds up to a fact about consciousness. . . .

Argument 5: From the Absence of Analysis

If proponents of reductive explanation are to have any hope of defeating the arguments above, they will have to give us some idea of how the existence of consciousness might be entailed by 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. But 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—and there is no such analysis to be had.

The only analysis of consciousness that seems even remotely tenable for these 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 that it make information available to later processes in a certain way, or whatever. But on the face of it, 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 the 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 becomes much easier, but it is a hollow achievement.

Functional analyses of consciousness can also be argued against on more specific grounds. For example, any functionally analyzed concept will have a degree of semantic indeterminancy. Does a mouse have beliefs? Do 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, and in any high-level functional concepts the boundaries will be vague. But compare: Does a mouse have conscious experience? Does a virus? These are not matters for stipulation. Either there is something that it is like to be a mouse or there is not, and it is not up to us to define the mouse's experiences into or out of existence. 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. This determinacy could not be derived from any functional analysis of the concepts in the vicinity of consciousness, as the functional concepts in the vicinity are all somewhat vague. If so, it follows that the notion of consciousness cannot be functionally analyzed.

Another objection is that the functional analysis collapses the important distinction . . . between the notions 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 in later processing and in the control of behavior. Awareness is a perfectly good concept, but it is quite distinct from the concept of conscious experience. The functionalist treatment collapses the two notions of consciousness and awareness into one, and therefore does not do justice to our conceptual system.

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 clearly inadequate. 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. It seems that the concept of consciousness is irreducible, being characterizable only in terms of concepts that themselves involve consciousness.

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 easy to see that properties such as life, learning, and so on can be analyzed as functional properties, 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.

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 completely fail 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 and of learning. 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.

---------- *end of required reading* ----------

2. The Failure of Reductive Explanation

The failure of consciousness to logically supervene on the physical tells us that no reductive explanation of consciousness can succeed. Given any account of the physical processes purported to underlie consciousness, there will always be a further question: Why are these processes accompanied by conscious experience? For most other phenomena, such a question is easily answered: the physical facts about those processes *entail* the existence of the phenomena. For a phenomenon such as life, for example, the physical facts imply that certain functions will be performed, and the performance of those functions is all we need to explain in order to explain life. But no such answer will suffice for consciousness.

Physical explanation is well suited to the explanation of *structure* and of *function*. Structural properties and functional properties can be straightforwardly entailed by a low-level physical story, and so are clearly apt for reductive explanation. And almost all the high-level phenomena that we need to explain ultimately come down to structure or function: think of the explanation of waterfalls, planets, digestion, reproduction, language. But the explanation of consciousness is not just a matter of explaining structure and function. Once we have explained all the physical structure in the vicinity of the brain, and we have explained how all the various brain functions are performed, there is a further sort of explanandum: consciousness itself. Why should all this structure and function give rise to experience? The story about the physical processes does not say.

We can put this in terms of the thought experiments given earlier. Any story about physical processes applies equally to me and to my zombie twin. It follows that nothing in that story says why, in my case, consciousness arises. Similarly, any story about physical processes applies equally to my inverted twin, who sees blue where I see red: it follows that nothing in that story says why my experience is of one variety rather than another. The very fact that it is logically possible that the physical facts could be the same while the facts about consciousness are different shows us that as Levine (1983) has put it, there is an *explanatory gap* between the physical level and conscious experience.

If this is right, 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 must be taken as brute. We might try to systematize and explain these brute facts in terms of some simple underlying pattern, but there will always remain an element here that is logically independent of the physical story. Perhaps we might get some kind of explanation by combining the underlying physical facts with certain further bridging principles that link the physical facts with consciousness, but this explanation will not be a reductive one.

The very need for explicit bridging principles shows us that consciousness is not being explained reductively but is being explained on its own terms.

Of course nothing I have said implies that physical facts are *irrelevant* to the explanation of consciousness. We can still expect physical accounts to play a significant role in a theory of consciousness, giving information about the physical *basis* of consciousness, for example, and perhaps yielding a detailed correspondence between various aspects of physical processing and aspects of conscious experience. Such accounts may be especially useful in helping to understand the *structure* of consciousness: the patterns of similarity and difference between experiences, the geometric structure of phenomenal fields, and so on. . . . But a physical account, alone, is not enough.

At this point, a number of objections naturally arise.

Objection 1: Are We Setting the Standards Too High?

Some might argue that explanation of *any* high-level phenomena will postulate "bridge laws" in addition to a low-level account, and that it is only with the aid of these bridges that the details of the high-level phenomena are derived. However, as the discussion in the last chapter suggests (and as is carefully argued by Horgan [1978]), in such cases the bridge laws are not further facts about the world. Rather, the connecting principles themselves are logically supervenient on the low-level facts. The extreme case of such a bridging principle is a supervenience conditional, which we have seen is usually a conceptual truth. Other more “localized” bridging principles, such as the link between molecular motion and heat, can at least be derived from the physical facts. For consciousness, by contrast, such bridging principles must be taken as primitive.

It is interesting to see how a typical high-level property—such as life, say—evades the arguments put forward in the case of consciousness. First, it is straightforwardly inconceivable that there could be a physical replica of a living creature that was not itself alive. Perhaps a problem might arise due to context-dependent properties (would a replica that forms randomly in a swamp be alive, or be human?), but fixing environmental facts eliminates 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 about its environment), one has enough material 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. As a general point, most high-level phenomena come down to matters of physical structure and function, and we have good reason to believe that structural and functional properties are logically supervenient on the physical.

Objection 2: Couldn't a Vitalist Have Said the Same Thing about Life?

All this notwithstanding, a common reaction to the sort of argument I have given is to reply that a vitalist about life might have said the same things. For example, 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. And a vitalist might have argued that life is a further fact, not explained by any account of the physical facts. But the vitalist would have been *wrong*. By analogy, might not the opponent of reductive explanation for consciousness also be wrong?

I think this reaction misplaces the source of vitalist objections. Vitalism was mostly driven by doubt about whether physical

mechanisms could perform all the complex *functions* associated with life: adaptive behavior, reproduction, and the like. At the time, very little was known about the enormous sophistication of biochemical mechanisms, so this sort of doubt was quite natural. But implicit in these very doubts is the conceptual point that when it comes to explaining life, it is the performance of various functions that needs to be explained. Indeed, it is notable that as physical explanation of the relevant functions gradually appeared, vitalist doubts mostly melted away. With consciousness, by contrast, the problem persists even when the various functions are explained.

Presented with a full physical account showing how physical processes perform the relevant functions, a reasonable vitalist would concede that life has been explained. There is not even *conceptual* room for the performance of these functions without life. Perhaps some ultrastrong vitalist would deny even this, claiming that something is left out by a functional account of life—the vital spirit, perhaps. But the obvious rejoinder is that unlike experience, the vital spirit is not something we have independent reason to believe in. Insofar as there was ever any reason to believe in it, it was as an explanatory construct—"We *must* have such a thing in order to be able to do such amazing stuff." But as an explanatory construct, the vital spirit can be eliminated when we find a better explanation of how the functions are performed. Conscious experience, by contrast, forces itself on one as an explanandum and cannot be eliminated so easily.

One reason a vitalist might think something is left out of a functional explanation of life is precisely that nothing in a physical account explains why there is something it is like to be alive. Perhaps some element of belief in a "vital spirit" was tied to the phenomena of one's inner life. Many have perceived a link between the concepts of life and experience, and even today it seems reasonable to say that one of the things that needs to be explained about life is the fact that many living creatures are conscious. But the existence of *this* sort of vitalist doubt is of no

comfort to the proponent of reductive explanation of consciousness, as it is a doubt that has never been overturned.

Objection 3: Is Conceivability a Guide to Possibility?

Philosophers are often suspicious of arguments that give a key role to conceivability, frequently responding that conceivability does not suffice for possibility. This is a subtle issue that I have discussed earlier and will discuss again: but here, the subtleties are not especially relevant. When it comes to matters of *explanation*, it is clear that conceivability is central. If on reflection we find it conceivable that all these physical processes could take place in the absence of consciousness, then no reductive explanation of consciousness will be satisfactory: the further question of why we exist and not zombies will always arise. Even if conceivability is tied to the limits of human capacity, explanation is tied to the limits of human capacity in a similar way.

Another way to put the point is to note that reductive explanation of a phenomenon in terms of the physical requires an *a priori* implication from the physical facts to the relevant high-level facts (logical supervenience according to primary intension, as I put it earlier). If such a connection does not hold, then we will always be able to raise the further question of why the physical processes give rise to consciousness. We have seen that in almost all domains, the right sort of connection holds, making reductive explanation possible; but it does not seem to hold for conscious experience. One can question whether *ontological* views such as materialism turn on these *a priori* links—I discuss that matter in the next chapter—but when it comes to reductive explanation, such links are crucial.

Objection 4: Isn't This a Collection of Circular Intuitions?

It might be further objected that the arguments I have given consist, at bottom, in a collection of 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 example, 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.

The "intuition" at work here is the very raison d'être of the problem of consciousness. The only consistent way to get around the intuitions is to deny the problem and the phenomenon altogether. One can always, at least when speaking "philosophically," deny the intuitions altogether, and deny that there is anything (apart from the performance of various functions) that needs explaining. But if one takes consciousness seriously, the conclusions for which I am arguing must follow.

Objection 5: Doesn't All Explanation Have to Stop Somewhere?

A final objection is that no explanation gives one something for nothing: all explanation has to stop somewhere. In explaining the motion of the planets, for example, one takes the laws of gravity and the existence of mass for granted. Perhaps we should simply take something for granted in this case, too? I am sympathetic with this point; I think we do have to take something for granted in explaining consciousness. But in doing so we inevitably move

beyond a *reductive* explanation. Indeed, this sort of analogy lends support to the nonreductive position I am advocating. We take the laws of physics for granted because they are *fundamental* laws. If we take a link between physical processes and conscious experience for granted, this suggests that the link should be taken as fundamental in the same way. . . .

5. The Appeal to a New Physics

Sometimes it is held that the key to the explanation of consciousness may lie in a new sort of physical theory. Perhaps, in arguing that consciousness is not entailed by the physics of our world, we have been tacitly assuming that the physics of our world is something like physics as we understand it today, consisting in an arrangement of particles and fields in the spatiotemporal manifold, undergoing complex processes of causation and evolution. An opponent might agree that nothing in *this* sort of physics entails the existence of consciousness, but argue that there might be a new kind of physical theory from which consciousness falls out as a consequence.

It is not easy to evaluate this claim in the absence of any detailed proposal. 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. The crucial question is: How could a theory that is recognizably a physical theory entail the existence of consciousness? If such a theory consists in a description of the structure and dynamics of fields, waves, particles, and the like, then all the usual problems will apply. And it is unclear that *any* sort of physical theory could be different enough from this to avoid the problems.

The trouble is that the basic elements of physical theories seem always to come down to two things: the structure and dynamics and physical processes. Different theories invoke different sorts of

structure. Newtonian physics invokes a Euclidean space-time; relativity theory invokes a non-Euclidean differential manifold; quantum theory invokes a Hilbert space for wave functions. And different theories invoke different kinds of dynamics within those structures: Newton's laws, the principles of relativity, the wave equations of quantum mechanics. But from structure and dynamics, we can only get more structure and dynamics. This allows the possibility of satisfying explanations of all sorts of high-level structural and functional properties, but conscious experience will remain untouched. No set of facts about physical structure and dynamics can add up to a fact about phenomenology.

Of course, there is a 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 will certainly not be a reductive one. For our purposes, it is best to take physics to be the fundamental science developed to explain observations of the external world. If this kind of physics entailed the facts about consciousness, without invoking consciousness itself in a crucial role, then consciousness would truly be reductively explained. For the reasons I have given, however, there is good reason to believe that no such reductive explanation is possible.

Almost all existing proposals concerning the use of physics to explain consciousness focus on the most puzzling part of physics, namely quantum mechanics. This is understandable: for physics to explain consciousness would take something extraordinary, and quantum mechanics is by far the most extraordinary part of contemporary physics. But in the end it does not seem to be extraordinary enough.

For example, Penrose (1994) suggests that the key to understanding consciousness may lie in a theory that reconciles quantum theory with the theory of general relativity. He suggests that gravitational effects not yet understood may be responsible for the collapse of the quantum wave function, leading to a non-algorithmic element in the laws of nature. Drawing on the ideas of Hameroff (1994), he suggests that human cognition may depend on quantum collapse in microtubules, which are protein structures found in the skeleton of a neuron. Indeed, Penrose and Hameroff suggest that quantum collapse in microtubules may be the physical basis of conscious experience.

These ideas are extremely speculative, but they could at least *conceivably* help to explain certain elements of human cognitive functioning. Penrose suggests that the nonalgorithmic element in collapse could explain certain aspects of our mathematical insight, which he believes goes beyond the capacity of any algorithmic system. Hameroff suggests that the collapse of a superposed wave function might help explain certain aspects of human decision making. But nothing here seems to help with the explanation of conscious experience. Why should quantum processes in microtubules give rise to consciousness? The question here is just as hard as the corresponding question about classical processes in a classical brain. When it comes to the problem of experience, nonalgorithmic and algorithmic processes are in the same boat.

Some have suggested that the *nonlocality* 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. But even if physics is nonlocal, it is hard to see how this should help in the explanation of consciousness. Even given a nonlocal physical process, it remains logically possible that the process could take place in the absence of consciousness. The explanatory gap is as wide as ever.

The most frequently noted connection between consciousness and quantum mechanics lies in the fact that on some interpretations of

the latter, measurement by a conscious observer is required to bring about the collapse of the wave function. On this sort of interpretation, consciousness plays a central role in the dynamics of the physical theory. These interpretations are highly controversial, but in any case it is notable that they do nothing to provide an explanation of consciousness. Rather, they simply assume the existence of consciousness, and use it to help explain certain physical phenomena. Theories of consciousness that exploit this relationship are occasionally put forward (e.g., Hodgson 1988; Stapp 1993), but they are certainly not reductive theories.

One cannot rule out the possibility that fundamental physical theories such as quantum mechanics will play a key role in a theory of consciousness. For example, perhaps consciousness will turn out to be associated with certain fundamental physical properties, or with certain configurations of those properties, or perhaps there will be a more subtle link. But all the same, there is little hope that this sort of theory will provide a wholly physical *explanation* of consciousness. When it comes to reductive explanation, physics-based theories are no better off than neurobiological and cognitive theories.

6. Evolutionary Explanation

Even those who take consciousness seriously are often drawn to the idea of an evolutionary explanation of consciousness. After all, consciousness is such a ubiquitous and central feature that it seems that it must have arisen during the evolutionary process for *a reason*. In particular, it is natural to suppose that it arose because there is some function that it serves that could not be achieved without it. If we could get a clear enough idea of the relevant function, then we would have some idea of why consciousness exists.

Unfortunately, this idea 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 must therefore find something wrong with the arguments at the start of this chapter, rather than raising the question here.

To see the point in a different way, 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 truths, metaphysical necessities, or natural laws—are constant over spacetime: if a physical replica of me had popped into existence a million years ago, it would have been just as conscious as I am. The connecting principles themselves are therefore independent of the evolutionary process. While evolution can be very useful in explaining why particular physical systems have evolved, it is irrelevant to the explanation of the bridging principles in virtue of which some of the systems are conscious.

7. Whither Reductive Explanation?

It is not uncommon for people to agree with critiques of specific reductive accounts, but to qualify this agreement: "Of course *that* doesn'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. The problem lies in the overall explanatory strategy. These models and theories are simply not the *sort* of thing that could explain 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, artificial life, and quantum mechanics will provide only more powerful functional explanations. This may make for some very interesting cognitive science, but the mystery of consciousness will not be removed.

Any account given in purely physical terms will suffer from the same problem. It will ultimately be given in terms of the structural and dynamical properties of physical processes, and no matter how sophisticated such an account is, it will yield only more structure and dynamics. While this is enough to handle most natural phenomena, the problem of consciousness goes beyond any problem about the explanation of structure and function, so a new sort of explanation is needed.

It might be supposed that there could eventually be a reductive explanatory technique that explained something other than structure and function, but it is very hard to see how this could be possible, given that the laws of physics are ultimately cast in terms of structure and dynamics. The existence of consciousness will always be a further fact relative to structural and dynamic facts, and so will always be unexplained by a physical 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 *reductive* explanation. The possibility of explaining consciousness nonreductively remains open. This would be a very different sort of explanation, requiring some

radical changes in the way we think about the structure of the world. But if we make these changes, the beginnings of a theory of consciousness may become visible in the distance.

REFERENCES

Block, N. 1978. “Troubles with functionalism”, in N. Block, ed. *Readings in the Philosophy of Psychology*. Cambridge, MA: Harvard University Press.

Block, N., 1995. “On a confusion about a function of consciousness.” *Behavioral & Brain Sciences* 18: 227-47.

Hameroff, S. R., 1994. “Quantum coherence in microtubules.” *Journal of Consciousness Studies* 1: 91-118.