David Chalmers, 'The hard problem of consciousness'

Excerpts from David Chalmers, 'The hard problem of consciousness', in *The Norton Introduction to Philosophy*, edited by Gideon Rosen, Alex Byrne, Joshua Cohen, and Seana Shiffrin (Norton, 2015).

Chalmers begins by asking why 'physical processing in the brain give[s] rise to a conscious inner life: consciousness of shapes, colors, sounds, emotions, and a stream of conscious thought, all experienced from the first-person point of view'. This is, he says, 'perhaps the most baffling problem in the science of the mind'. Chalmers outlines his article as follows:

> 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. In the second half of the paper, I argue that if we move to a new kind of explanation that does not try to reduce consciousness to something it is not, a naturalistic account of consciousness can be given.

To isolate the 'truly hard part' he distinguishes the 'easy' problems from the 'hard' ones.

> 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 has put it, there is *something it is like* to be a conscious organism.[1] 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

[1] Chalmers is referring to Nagel's 'What is it like to be a bat?', which we read earlier.

> 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."

Having explained the difference between the easy and hard problems, Chalmers now turns to the question of why the 'easy problems' really are easy, and why the 'hard problem' really is 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 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.

Throughout the higher-level sciences, reductive explanation—explanation that explains a high-level phenomenon wholly in terms of lower-level phenomena—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*[2]…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.

[2] This term was coined by the American philosopher Joseph Levine, in his 1983 paper 'Materialism and qualia: the explanatory gap'.

> 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.

Chalmers now connects the explanatory gap with 'zombies' in the philosophers' sense:

> The hard problem of consciousness arises for any physical explanation of consciousness. For any physical process we specify there will be an unanswered question: why should this process give rise to experience?
>
> One way to see this point is via a philosophical thought-experiment: that of a philosophical zombie. A philosophical zombie is a being that is atom-for-atom identical to a conscious being such as you and me, but it is not conscious. Unlike the zombies found in Hollywood movies, philosophical zombies look just like a normal humans from the outside, and their behavior is indistinguishable from that of a conscious being. But on the inside, all is dark. There is nothing it is like to be a zombie.
>
> There is little reason to think that philosophical zombies really exist. But what matters for our purposes is simply that the idea is coherent. There is no internal contradiction in the idea of a zombie, the way that there is an internal contradiction in the idea of a round square. I may believe that you are not a zombie, but I cannot rule out the hypothesis that you are a zombie by a priori reasoning alone.
>
> The hard problem of consciousness might then be put as the problem: why are we not zombies? In our world, in fact, there is consciousness. But everything in physics and in neuroscience seems to be compatible with the hypothesis that we are zombies. If that is right, then physics and neuroscience alone cannot explain why we are not zombies. More generally, it appears that no purely physical explanation can explain why we are not zombies. If so, no purely physical explanation can solve the hard problem of consciousness.
>
> We can even use this sort of reasoning to generate an argument against materialism, the thesis that our world is wholly physical. To explain materialism, we can use the metaphor of God creating the world. If materialism is true, then God simply needed to create microphysical entities

such as atoms and fields, and arrange them in the right way: then everything else, such as cells and organisms and tables, followed automatically.

But zombies suggest that materialism must be false. To see this, note that because there is no contradiction in the idea of a zombie, it seems that it would be within God's powers to create a zombie world: a world that is physically identical to ours, but without consciousness. If this is right, then even after God ensured that all the physical truths about our world obtained, the truths about consciousness did not automatically follow. After creating everything in physics, God had to do more work to put consciousness into the world. This suggests that consciousness is something over and above the physical, and that materialism is false.

Of course God here is a metaphor, but the idea can also be put in terms of the philosophers' idea of a possible world. For example, there may be no antigravity machines in the actual world, but there is no contradiction in the idea (one can tell coherent science fiction about antigravity), so there is at least a possible world in which there is antigravity. Likewise, even if there are no zombies in the actual world, there is at least a possible world in which there are zombies. And if there is a possible world in which there are physical processes just like those in our world but no consciousness, then consciousness does not follow from those processes with absolute necessity. It follows that materialism is false.

We might put the underlying problem as follows. Physical explanation is ultimately cast entirely in terms of microphysical structure and dynamics. This sort of explanation is well-suited to explaining macroscopic structure and dynamics. For problems such as the problem of learning or the problem of life, this is good enough, as in these cases macroscopic structure and dynamics were all that needed explaining. But we have seen that in the case of consciousness, structure and dynamics is not all that needs explaining: we also need to explain why macroscopic structure and dynamics is accompanied by consciousness. And here, physical explanation has nothing to say: structure and dynamics adds up only to more structure and dynamics. So consciousness cannot be wholly explained in physical terms.

If all this is right, then although consciousness may be associated with physical processing in systems such as brains, it is not reducible to that processing. Any *reductive* explanation of consciousness, in purely physical terms, must fail. No matter what sort of physical processes we might invoke, we find an explanatory gap between those processes and consciousness.

At this point some are tempted to give up, holding that we will never have a theory of conscious experience. 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 ex- planation 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, the view that the mind is not wholly physical, as it postulates basic mental properties over and above the properties invoked by physics. But it is a version of dualism that is 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 properties connected by fundamental laws. It expands the class of primitive properties, 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*.

Chalmers concludes:

> 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.