Zombies and the Explanatory Gap

Section 4 of "The Hard Problem of Consciousness" in *The Norton Introduction to Philosophy*

by David Chalmers (2018)

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 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.