ZAPPING THE ZOMBIES

Robert Kirk

In the philosophy of mind, zombies often make an appearance. It seems we can conceive of zombies — beings physically exactly like ourselves but lacking conscious experience. There may not actually be any zombies, of course. But the suggestion that they could exist does at least seem to make sense. Or does it? Robert Kirk investigates.

1. Zombies

In Caribbean folklore zombies are corpses magically caused to walk about and work. They are like marionettes except that they are manipulated by magicians using paranormal forces, not normal people using strings. If horror films are anything to go by, such creatures would look and behave very differently from us, and it is obvious that they would not be conscious — any more than marionettes are conscious. The philosophical zombies I hope to interest you in would be very different. They would both look and behave exactly like us. They would not be corpses but fully functioning human bodies, physically just like human beings down to the tiniest neurophysiological details. Yet *by definition* they would have no conscious experiences at all: there is supposed to be 'nothing it is like' to be a zombie.

Not many people would say there actually are any zombies, but some claim there could have been such creatures. It is widely accepted that if zombies are so much as possible, then there is more to human consciousness than can be accounted for in purely physical terms. If they are possible, we have non-physical components whose presence explains the fact that there is something it is like to be one of us; in which case the world is not a purely physical system. That is one reason why the idea of zombies, weird as it is, is worth bothering about. I think an even more important reason is that in spite of its appeal it involves a grotesquely mistaken conception of consciousness. That conception is a huge obstacle to a true

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

understanding of consciousness and responsible for a lot of desperate and confused reasoning. (I think it underlies much of the opposition to 'functionalist' approaches to the nature of consciousness.) Wittgenstein, Ryle, and Dennett have all offered arguments which would imply that zombies are not conceivable. The trouble is that their arguments are widely thought to put too much weight on behaviour, and not to take sufficient account of what happens inside the organism. The argument to be set out here is not open to that objection. I think it also has some appeal to the imagination, so that it is a useful device for loosening the imaginative grip of the zombie idea.

2. The epiphenomenal-qualia conception of consciousness

In the nineteenth century scientists began to think there were good reasons to suppose that *every physical effect has a physical cause*. The developing science of neurophysiology looked set to provide explanations of the whole range of behaviour in terms of physical processes inside the body, interacting with the environment via stimulation of the sense organs. But if that is right, what is to be said about thoughts and feelings? One suggestion was that they too were just physical processes. That struck many as absurd, especially for the case of consciousness. T. H. Huxley and others had an interesting alternative suggestion. They continued to insist that every effect has a physical cause — that the physical universe is 'closed under causation' — since that seemed to be the irresistible message of science; but they did not understand how consciousness could be purely physical. They concluded that conscious experiences are a *non-physical by-product* of the workings of the brain; and because the causal closure of the physical entailed that nothing non-physical could affect anything physical, they had to maintain also that conscious experiences had no effects on the physical world. In their opinion human beings are 'conscious automata'. The world affects our sense organs, which in turn cause conscious experiences; but conscious experiences don't affect it. That view is *epiphenomenalism*. The supposed non-physical components

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

of experience according to that view are usually called *qualia* (the singular is *quale*); I will call them 'epiphenomenal qualia': 'e-qualia' for short.

As G. F. Stout pointed out long ago, if epiphenomenalism were true, it ought to be credible that the entire physical history of the universe should have been 'just the same as it is if there were not and never had been any experiencing individuals. Human bodies would still have gone through the motions of making and using bridges, telephones and telegraphs, of writing and reading books, of speaking in Parliament, of arguing about materialism, and so on' (*Mind and Matter,* 1931, pp.138f.). Although Stout didn't use the word 'zombie', he was describing a zombie world: a world where all physical effects had physical causes, and whose inhabitants were exactly like us in all physical respects but had no e-qualia. For convenience (and with apologies for an unpleasant piece of jargon) I will refer to the conception of human consciousness implied by epiphenomenalism as the 'e-qualia conception'.

Epiphenomenalism, then, entails that a zombie world is at least *conceivable,* in the sense that it involves no contradiction or other incoherence. But the converse is not true. To hold that a zombie world is conceivable does not commit you to epiphenomenalism even if you are a dualist, since you could maintain (for example) that in the world as it actually is, the non-physical items involved in conscious experience have, unlike e-qualia, physical effects. However, if you hold that a zombie world is so much as conceivable, then although you are not committed to epiphenomenalism, you are still committed to the conceivability of the e-qualia model of consciousness, or so I would argue. That is, you are committed to there being no incoherence in the idea of a world satisfying the following conditions:

E1 Every physical effect has a physical cause. (The causal closure of the physical.)

E2 Human beings are physical systems with epiphenomenal qualia. These e-qualia are special non-physical properties which make it the case

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

that we are phenomenally conscious: that there is 'something it is like' to have experiences.

E3 E-qualia are caused by physical processes in our bodies but have no physical effects; they could be stripped off without disturbing the physical world.

E4 Human beings consist of nothing else but functioning bodies and their e-qualia.

If you think zombies are conceivable — that their description involves no contradiction or other incoherence — then obviously you are committed to the conceivability of E1; for zombies are supposed to be in all physical respects just as we should be if the physical world were indeed closed under causation. You are also obviously committed to there being a certain crucial *something* which is non-physical, makes the difference between a zombie and a conscious individual, yet has no physical effects. Plausibly, therefore, you are committed to the conceivability of e-qualia as explained in E2, E3, and E4: you are committed to:

(A) The conceivability of zombies entails the conceivability of the e-qualia conception.

It would take more than what I have just said to clinch the argument for (A). I think it can be done by showing that if you accept the conceivability of zombies, you cannot deny that it is also conceivable that the world as you believe it to be could be *transformed* into a world where conditions E1-E4 were satisfied; in which case you must accept that the e-qualia conception is conceivable. But it would be too distracting to set out the argument here, so in this discussion I will just assume (A).

Now for the main argument, which I think will show that:

(B) The e-qualia conception of consciousness is incoherent.

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

If that is right, then, given (A), it will follow that:

(C) The zombie idea too is incoherent, and zombies are impossible for that reason.

(The argument has the form: If p then q; but not-q; therefore not-p.) Even if (A) is mistaken, and the friends of zombies can somehow escape commitment to the e-qualia model, the inconceivability of the e-qualia model is still highly significant. For that model is presupposed not only by most epiphenomenalists but by others, including many physicalists.

3. The e-qualia conception and causation

According to clause E3, e-qualia have no physical effects. That view is characteristic of epiphenomenalism — and widely held to be its fatal flaw. Recall that e-qualia are supposed to be the very things that ensure there is *something it is like* for us to have conscious experiences, for example perceptual experiences such as seeing colours or smelling scents. It certainly seems that we are continually being affected by such experiences in ways which have effects on our behaviour. For example, I may prefer the particular character of the experience I have when I taste this wine to the one I have when I taste that wine, and buy another bottle of this wine. Now, epiphenomenalists are well aware of this particular objection and have a reply to it. They concede that it *seems* that the character of my experiences when tasting the wine contributes to causing the physical activities involved in buying the bottle, but they insist that the reality is different. What happens in reality, they claim, is that only the purely *physical* processes involved in tasting the wine are causal factors in my behaviour. It only seems that the experience itself was a causal factor because that experience — strictly, the e-quale which occurred when I considered the particular character of the taste — was itself caused by the same physical processes: first the physical processes caused the experience, then they caused the behaviour, and the result is that it seemed to me that the experience caused the behaviour when it didn't. In support of

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

this account epiphenomenalists can point out that most of us know nothing about the underlying physical processes, and are therefore easily misled.

Does that manoeuvre allow the e-qualia model to escape the objection? There are several further difficulties. One is that it is hard to see how we could so much as *refer* to our experiences if they had no effects on the physical activities involved in thinking and talking about them; but I will not pursue that topic. The difficulty I will focus on arises in connection with activities such as *thinking about, attending to,* and *comparing* the qualities of our conscious experiences.

The e-qualia conception entails that those qualities are e-qualia. For by clause E2, it is e-qualia which ensure that there are conscious experiences; from which it follows that the e-qualia conception entails that zombies lack e-qualia; and it follows from clause E4 that e-qualia are the *only* things zombies lack. So according to the e-qualia model, the qualities that we think about, attend to, and compare can only be e-qualia. And the trouble is that the model seems to provide no way for us to be capable of such thinking about, attending to, or comparing of e-qualia.

4. E-qualia and causation

To see why, we must first notice a crucial consequence of the e-qualia model: *e-qualia do not have effects on other e-qualia.* Suppose for example I have an e-quale E associated with smelling eucalyptus, and E was caused by something or other. By clause E3, all e-qualia are caused to occur by physical events, so what caused my eucalyptus e-quale must have been some physical event. But by clause E2, e-qualia are non-physical, so the cause could not have been another e-quale. Since the particular example of smelling eucalyptus was quite arbitrary, it follows that, quite generally, e-qualia do not have effects on other e-qualia. (You might wonder whether E could have been caused by another e-quale as well as by something physical: an instance of 'overdetermination'. But that suggestion involves a contradiction. Keep in mind that we are considering (what are supposed to be) actual facts. If

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

all e-qualia are actually caused to occur by physical events, there is no work for e-qualia to do: they make no difference to the course of events, and so are not causes at all.)

E-qualia, then, are completely inert. By E3, they don't cause physical events, and by the last paragraph they don't cause non-physical events either. That has serious consequences for the e-qualia model, as becomes clear when we consider what must be involved in thinking about, attending to, and comparing e-qualia. These are not simple activities. As everyone agrees, they involve complicated cognitive processing, for example conceptualization and the storing and retrieving of information. Those processes in turn necessarily involve the causation of changes. Storing information involves the causation of 'traces' or other effects; retrieving information involves effects on the organism's current capacities; thinking and conceptualization similarly involve changes causing other changes. Since e-qualia are causally inert, they cannot perform such activities.

Now, we noticed earlier that epiphenomenalists say that although it *seems* that our behaviour is affected by the qualities of our experiences, that is a mistake: in reality they do not affect the physical world. Could defenders of the e-qualia model similarly maintain that we do not really think about, attend to, or compare e-qualia? Can they say it is all an illusion? Clearly not. That would imply that they could not think about or attend to the items they spend so much time talking about! So they are bound to concede that we really engage in those activities. Since, as we have just seen, they cannot claim those activities are performed by the e-qualia themselves, they are forced to assign the work to our bodies. For by clause E4 there is no more to us than functioning bodies and e-qualia. I will now try to make clear that our bodies cannot do that work either.

5. My zombie twin's sole-pictures

Assume for argument's sake that the e-qualia model is conceivable, so that the zombie idea is conceivable too; and suppose I have a zombie twin called Zob. He is an exact physical duplicate of myself in a zombie world which is an

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

exact physical duplicate of this world and subject to causal closure; but, being a zombie, he is not conscious. By clause E4 of the e-qualia conception the only difference between him and me is that he has no e-qualia. But that conception entails that if the natural laws of Zob's world were suddenly to change in such a way that he acquired suitable e-qualia, the result would be a fully conscious being — a complete duplicate of me. I am not saying *Zob* would be conscious, since he is just a zombie; but on the e-qualia model the result must be *a* conscious being. Could this conscious being think about, attend to, or compare e-qualia?

We saw in the last section that the e-qualia conception entails that those activities must be carried out by purely *physical* processes. That was because e-qualia are completely inert: cannot have any effects on anything. The crucial question now is this: given that Zob somehow acquires suitable e-qualia, what can ensure that the cognitive processes in his brain and body are in any way *about* those e-qualia, or constitute *attending to* or *comparing* e-qualia?

The point of that question can be brought out by means of another. Why should Zob's acquisition of e-qualia result in an individual who stands in any of those relations with his experiences, when the presence of moving pictures on the soles of his feet would not result in anyone's standing in such a relation to those pictures? A silly question? I don't think so.

Since Zob is my zombie twin, his brain processes mirror mine. So whatever neural processes cause my visual e-qualia according to the e-qualia model are mirrored in Zob's brain. Suppose now that by an odd change in the natural laws of his world, those same physical visual processes in Zob cause sequences of constantly changing pictures on the soles of his feet. The coloured patterns on his soles mirror those neural processes just as my e-qualia are typically supposed to mirror similar processes in my brain. By 'mirror' I mean that there are one-to-one correspondences between the significant features of the two sets of processes, and one-to-one correspondences between the relations holding between those features. (In other words the two sets of processes are isomorphic.) If I

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

were magically transported to Zob's world, then, and able to view his sole-pictures, they would strike me as an accurate record of my own ongoing visual experiences. Now, could the fact that these sole-pictures were caused by and mirrored some of the cognitive processes in Zob's body ensure that his continuing cognitive processes were about them or constituted someone's attending to or comparing them? Obviously not. He never even notices his sole-pictures — I know that because I am telling the story — so his cognitive processes cannot be about them, or constitute anyone's being in any way acquainted with them. On that, at least, there will be general agreement.

6. Zob's e-qualia won't make anyone conscious

It follows that if Zob's acquiring a suitable lot of e-qualia results in a fully conscious subject coming into existence, there must be a relevant difference between sole-pictures and e-qualia. Of course there are some differences. Notably, Zob's sole-pictures are physical, and visible to anyone who gets into a position to see them; while his e-qualia are non-physical and therefore invisible. But do those differences bear on the question whether the cognitive processing in Zob's body makes anyone in any way acquainted with his e-qualia?

Imagine the sole-pictures themselves were suddenly to *become* non-physical, retaining as many of their original properties as possible. The suggestion is scarcely intelligible, of course, but it helps to bring out the irrelevance of physicality and non-physicality. Given that Zob's cognitive processing is not about his sole-pictures as things supposedly are, why should their becoming non-physical — or, less obscurely, their being superseded by non-physical items — make any relevant difference? I know of no reason why a thing's merely being non-physical should have any connection with consciousness at all, let alone with aboutness. It seems, then, that the non-physicality of e-qualia cannot make them relevantly different from Zob's sole-pictures.

Is the *invisibility* of e-qualia a relevant difference? If anything, the opposite seems to be the case. After all, what ensures

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

that Zob's cognitive processing has nothing to do with his sole-pictures is that (as the story has it) he never *sees* them: they never have *effects* on those processes. But in principle he *could* see them; and if he did, then some of his cognitive processes would be about them in some respect or other. Since e-qualia are not just non-physical, and therefore invisible, but defined to have no effects at all, they would be even less accessible than sole-pictures. Their difference from sole-pictures in that respect is clearly not relevant.

Are there any other possibly relevant differences? Part of the trouble is that the e-qualia conception has little to say about the e-qualia themselves. They are supposed to be properties which ensure that their possessor is 'phenomenally conscious' — that there is something it is like — and they are supposed to be caused by physical events but causally inert themselves. They are also usually supposed to mirror certain (physical) perceptual processes. Beyond that, the e-qualia conception is silent.

One further thought: e-qualia may be said to *represent* those aspects of the world of which we are conscious. But what could be supposed to account for that, if it were correct? The only plausible suggestions would be (a) that they were systematically *caused* by those aspects of the world; or (b) that they *mirrored* those aspects of the world. But the same is true of Zob's sole-pictures. If e-qualia represent aspects of the world, so do his sole-pictures. So in that respect there is no relevant difference.

If there is some relevant difference between e-qualia and Zob's sole-pictures which I have overlooked, no doubt defenders of the e-qualia conception will correct me. But on the basis of what is said in current discussions of these matters we seem justified in concluding that there are no such differences. In that case the result of Zob's acquiring a suitable set of e-qualia will *not* be that anyone can think about, attend to, or compare his e-qualia. The e-qualia model allows people no more access to their e-qualia than Zob has to his sole-pictures: that is, none.

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

The e-qualia conception of consciousness therefore cannot do what it sets out to do. It cannot account for (among other things) our capacities to think about, attend to, and compare the qualities of our experiences. According to that conception human consciousness has just two components. One is the functioning body, which does all the necessary cognitive processing; the other is a complex of e-qualia, which are supposed to ensure that there is something it is like. But it lacks the resources to enable the cognitive processing to be *about* the e-qualia, or to result in any sort of acquaintance with them. So the e-qualia conception is not merely odd or factually mistaken: it is fundamentally incapable of working. It is incoherent.

Recall that in section 2 I sketched reasons for accepting the following thesis:

(A) The conceivability of zombies entails the conceivability of the e-qualia conception of consciousness.

From the reasoning just outlined we get the other main thesis I picked out:

(B) The e-qualia conception of consciousness is incoherent.

From (A) and (B) it follows that:

(C) The zombie idea too is incoherent, and zombies are impossible for that reason.

7. Conclusion

When you first come across the idea of zombies it seems to make perfectly good sense. It is easy to suppose, with T. H. Huxley and other epiphenomenalists, that our experiences are non-physical *additions* to a fully functioning physical world, extras which have no effects on that world and are related to it at most by being caused by, and (possibly) mirroring,

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press

certain processes in it. If you accept that model, then you are committed to its being conceivable that those non-physical extras should be stripped off like a jacket, leaving the physical world churning on exactly as before. You are therefore also committed to the conceivability of zombies. But if the sole-pictures argument is sound that model cannot possibly work. Of course the argument may be unsound; certainly there are possible objections that I have not discussed here. But if the argument is sound, consciousness is not the sort of thing that could have been stripped off in that way, leaving our bodies continuing to function unchanged. That whole conception of consciousness is fundamentally mistaken.

Robert Kirk is Emeritus Professor of Philosophy at the Department of Philosophy, University of Nottingham. His most recent book is Zombies and Consciousness *(Oxford: OUP, 2005)*

https://doi.org/10.1017/S1477175600001536 Published online by Cambridge University Press