Journal of Consciousness Studies
controversies in science and the humanities

Conversations with Zombies

Todd C. Moody
Department of Philosophy
St. Joseph's University
5600 City Avenue
Philadelphia PA 19119
USA

Email: tmoody@sjuphil.sju.edu

***Journal of Consciousness Studies*, 1 (2), 1994, pp. 196-200.**

The problem of 'conscious inessentialism' is examined in the literature, and an argument is presented that the presence of consciousness is indeed marked by a behavioural difference, but that this should be looked for at the _cultural_ level of speech communities.

The 'zombie problem' is the problem of consciousness, stated in a particularly provocative way. Given _any_ functional description of cognition, as detailed and complete as one can imagine, it will still make sense to suppose that there could be insentient beings that exemplify that description. That is, it is possible that there could be a behaviourally indiscernible but insentient simulacrum of a human cognizer: a zombie. This is so because the best functional description can only map inputs onto behaviours by means of computations of some sort. That certain computations are associated with consciousness is at most a contingent fact about them. The problem is reiterated at the level of any possible physical description of cognition as well. In this case, the intervening processes between inputs and behaviour will be of a causal, rather than formal, sort.[1] Nevertheless, the link between those processes and consciousness is still contingent. As long as the link between publicly observable states of any sort and consciousness is contingent, zombies are a possibility. The zombie problem is a variation on the 'other minds' problem, but I hope to show that it is not an idle variation. It offers, I think, a vivid way of conceptualizing the philosophical questions about consciousness. Suppose there is a world much like our own, except for one detail: the people of this world are insentient. They engage in complex behaviours very similar to ours, including speech, but these behaviours are not accompanied by conscious experience of any sort. I shall refer to these beings as zombies.[2] This scenario, though surprising, is a possibility suggested by a theory recently referred to by Owen Flanagan as 'conscious inessentialism', which is defined as follows:

> the dominant philosophical theory of mind, _computational functionalism_ was (and still is) committed to the view of _conscious inessentialism_. This is the view that for any mental activity M performed in any cognitive domain D, even if _we_ do M with conscious accompaniments, M can in principle be done without these conscious accompaniments. (Flanagan, 1991)

Conscious inessentialism clearly entails that any given behaviour could also occur without conscious accompaniments. The only reason why one would suppose that certain behaviours do require conscious accompaniments is that the behaviours in question appear to require mental activity of some sort. Since conscious inessentialism tells us that no mental activity requires conscious accompaniments, it follows that no overt behaviour requires them either. So if conscious inessentialism is true, zombies are possible. Indeed, if conscious inessentialism is true, it is quite possible for an entire world of zombies to evolve, which is the premise of the current thought experiment. It is behaviours, after all, and not subjective states, that are subjected to evolutionary selection pressures. If those behaviours do not require consciousness, then evolution is indifferent to it. That the zombie problem may have significant metaphysical implications is concluded by Robert Kirk in a paper on the topic: 'it is hard to see how any intelligible version of Materialism could be reconciled with the logical possibility of Zombies, given that we are sentient'. (Kirk, 1974)

Is conscious inessentialism true? One argument for conscious inessentialism was intimated above: the most that we can ever hope to establish by empirical means is the regular correlation of observable states of some sort with consciousness. Such a correlation warrants only a conclusion of a contingent relation. Another source of support for conscious inessentialism comes from cognitive psychology. Increasingly, scientists are finding that what happens in consciousness is not essential for understanding mental functioning. We recognize each other, solve problems, use language, and although all these things have 'conscious accompaniments' it seems that the real work is not done consciously at all. In short, cognitive science is drifting towards a kind of epiphenomenalism. The artificial intelligence research programme is a part of this drift. To quote Flanagan again:

> Whereas most skeptics of strong artificial intelligence press worries that machines cannot be given consciousness, computational functionalism can be read as making this objection irrelevant. Mind does not require consciousness. (Flanagan, 1991)

The same point is made by John Searle, commenting on the intelligibility of the distinction between conscious and unconscious robots.

> as far as the ontology of consciousness is concerned, behavior is simply irrelevant. We could have _identical behavior_ in two different systems, one of which is conscious and the other totally unconscious. (Searle, 1992)

If conscious inessentialism is true, then it would presumably be impossible for us to tell whether visitors from another world are zombies. After all, if there is no necessary behavioural difference between them and us, as conscious inessentialism requires, there would be no identifiable mark of zombiehood. This at least is what appears to follow from the thesis. I shall argue that the 'mark of zombiehood' will be found not at the level of individuals but at the level of speech communities.

Let us begin by looking at the language of the zombies, and the sorts of things they might say with it. Let us suppose that they speak zombie- English, a language that looks much like our own English language. Certain words of zombie-English would have to have meanings somewhat different from their English counterparts. For example, the word 'understand' in English refers not only to what sorts of performances a person is capable of, given certain inputs and outputs, but also to a particular kind of conscious experience. I use this example because it is so familiar from Searle's Chinese Room argument. To understand Chinese is more than to be able to produce passable answers to questions. There is something it is like to understand Chinese, or anything else. Understanding has a phenomenology.

For a zombie, however, it is not like anything to understand Chinese, or anything else, because it is not like anything to do anything. So when a zombie uses the word 'understand' we must understand that he or she is not making any reference to any sort of conscious experience. To distinguish zombie-English words of this ilk from their English counterparts, I shall use the superscript[z]. Thus, we can say that zombies understand[z] many of the same things that we understand.

We can imagine being visited by the beings from zombie-Earth. Our natural science might look a great deal like theirs, and our mathematics would likely be very similar. That is, their beliefs[z] in these domains could easily be supposed to be very close to our beliefs. But there would also be some interesting differences, especially in the domain of philosophy.

Consider, for example, the phenomenon of dreaming. Could there be a cognate concept in zombie-English? How might we explain dreaming to them? We could say that dreams are things that we experience while asleep, but the zombies would not be able to make sense[z] of this. For them, the word 'experience[z]' can have at most a behavioural meaning. Indeed, zombies would understand[z] sleep[z] itself quite differently from the way we do. The word 'conscious[z]' could only mean responsive to the environment.

Still, we can imagine that it happens on the zombie world that upon waking[z] from sleep[z], one finds[z] that one has acquired apparent memories[z] that are not consistent with the rest of one's beliefs[z] and memories[z], and so forth. The zombies might refer to this coming-to-seem[z]-to-remember[z] as dreaming[z]. But that is not the end of the matter, as we shall presently see.

If there are zombie philosophers, they would be able to make no sense[z] at all of the other minds problem. They might, of course, be able to ponder[z] an 'other zombies' problem: how do we know[z] that there are not other beings whose experience[z] is accompanied by a quality that we cannot fathom[z]? But it is unlikely that this problem would occur to them, until they met us. At that point, the zombie philosophers might do a great deal of thinking[z] about the things that we try to explain to them. They would be especially puzzled[z] by our human philosophical literature about dreaming, especially the debate between those who argue that dreams are experiences and those who, like Dennett, argue that they can be explained as coming-to-seem-to-remember.[3] To the zombie philosophers, Dennett's position would be clear enough. What would be mysterious is the opposing (and, in fact, more 'common sense' to us) view, that dreams are experiences. So, even though zombies might dream[z], their concept[z] of dreaming[z] would not be philosophically problematic in the way that ours is.

Zombie philosophers would be persistently baffled[z] by the fact that we talk about 'consciousness' as if it were something more than being the subject[z] of experiences[z]. What more could it be? Would they be able to make sense[z] of the inverted spectrum problem? It is hard to see how they could. This is an interesting example, since even those human philosophers who argue that it is a pseudo-problem are nevertheless able to understand it.

The inverted spectrum problem would not work in zombie philosophy because it depends upon a concept of 'internal seeing'. This concept is found not only in academic (human) philosophy, but also in imaginative literature and film. In the 'Terminator' science-fiction films, for example, we are occasionally given a 'robot's-eye view' of the world, in which the bottom portion of the screen is filled up with various readouts: distance to target, velocity and the like. As human viewers of the film this device makes sense to us, because we can conceive of a kind of 'internal seeing' in which we can look at the readouts while we also look at the world. This imaginative device could make no sense[z] at all to zombies, because the idea of 'internally seen' readouts has no zombie analogue or purpose.[4]

Philosophers - human, earthbound philosophers - have argued that the other minds problem cannot be solved except by analogy, and that there is no empirical content to the notion of a zombie. That is, they argue that because there is no behavioural (and therefore observable) 'mark of zombiehood', it follows that the concept has no real content. But I hope that I have shown that while it is true that zombies who grew up in our midst might become glib in the use of our language, including our philosophical talk about consciousness and dreams, a world of zombies could not _originate_ these exact concepts as they are played out in philosophical discourse and imaginative idea-play, such as science fiction. Their discourse would have gaps in it (from our perspective), and concepts from our discourse (philosophical and imaginative) would be permanently untranslatable into theirs. This is important, because it suggests a qualification to conscious inessentialism. Even though the activities of talking about the philosophical dream problem or internal seeing do not require consciousness, the _emergence of those concepts in a language community_ does. This means that at the level of culture there are _necessary behavioural differences_ between zombies and non- zombies, because those differences are the result of the differences in the conceptual vocabularies available to each culture. At the level of _culture_, conscious inessentialism is false.

What is most interesting is the fact the zombie scientists would have to regard consciousness (not consciousness[z]) as something beyond the scope of their science. They would be forced to conclude[z] that consciousness is not consciousness[z]. But their science is methodologically just like ours. Suppose that human scientists were to develop what they took to be the complete scientific explanation of consciousness and deliver it to the zombie scientists, saying: 'Here is the full explanation of human consciousness. We hope it answers your questions.' It wouldn't, though. No matter how replete a scientific explanation of consciousness we might present to the zombie scientists, they would still have no inkling[z] of the explanandum. This is another way of stating Nagel's point that the scientific worldview explicitly excludes the subjective. (Nagel, 1986, Ch. 2).

That the zombies are different from us is a fact discoverable not by natural science but by a kind of hermeneutic analysis of the sorts of things that we talk about and what we have to say about them.[5] The mere fact that the zombies (as we would later recognize them to be) do not philosophize about internal seeing in the way that we do would not entail that they are zombies. That we could not explain that problem to them, however, would raise suspicions. Further analysis would clarify the conceptual gaps. But do these gaps cast a shadow of doubt upon materialism, as Robert Kirk claimed? Nagel claims that they do not:

> The fact that mental states are not physical states because they can't be objectively described in the way that physical states can doesn't mean that they must be states of something different. The falsity of physicalism does not require nonphysical substances. It requires only that things be true of conscious beings that cannot, because of their subjective character, be reduced to physical terms. (Nagel, 1986, p. 29)

These facts lead Nagel to a dual aspect theory, although he concedes that such a theory is ‘largely hand waving’ (_ibid._, p. 30). That is, reality has those aspects that can be encompassed within natural scientific theory and those aspects that, in principle, cannot. The conscious aspect of reality can neither be reductionistically eliminated nor explained by natural science. This means that there are not necessarily any discoverable physical differences between zombies and us that would explain the phenomenological difference. This counts against materialism because these phenomenological differences are perfectly real but are not part of the ‘natural order’, as it is materialistically construed. They cast a shadow, if you will, into the natural order in virtue of our ability to talk about them. Zombies may be able to ape our consciousness-talk, but they cannot _originate_ it with any hope of getting it right.

My own view is that this radical incompleteness of natural science with respect to consciousness entails, at the minimum, an equally radical agnosticism about the ontology of minds and persons. It means that we are not in a position to insist that materialism is true, and that therefore nonmaterialistic hypotheses and research programmes cannot be rejected a priori. The appeal of a dual aspect theory is that it avoids the difficulties of ontological dualism, but it is indeed mostly hand waving. It does not really explain why it should be that the stuff of the world has irreducibly distinct categories of properties. As I see it, dual aspect theory is largely an attempt to disguise the incompleteness of materialism. It is steadfastly materialistic at the level of ‘substance’ and quarantines the problems of dualism to the level of ‘properties’. Ontological agnosticism is more candid.

This line of thinking has some interesting corollaries. Zombies are, in relation to us, in the same predicament that most of us are in relation to those mystics who report back to us their experiences of what is sometimes called superconsciousness. We can ape what they say, if we want to, but we don’t really know what we are talking about. This difficulty is sometimes referred to in the mystical literature as ‘ineffability’, but the mystics understand each other, just as human non-zombies do.[6]

There is a literature on the question as to whether the things that mystics say count as evidence for the reality of a transcendent order of some sort. Given the absence of an independent way to verify their statements this presents grave difficulties. Nevertheless, we can easily imagine the parallel case of the zombie philosophers wondering[z] whether our consciousness-talk is evidence of something other than mere conciousness[z]. We can understand that they might be sceptical[z], even though to us there is nothing more real than consciousness.

Consider the possibility that a few zombies might discover a discipline that, after considerable practice, turns them into non-zombies, like us. It would presumably be very difficult to convince other zombies that such a discipline has any point, and it would be quite easy for the zombies to dismiss the phenomenon as marginal or pathological. The zombie scenario does not prove the ‘validity’ of mystical experience, whatever that would mean, but it does entail that such experience cannot be dismissed on the grounds of its radical unfamiliarity to the rest of us. We might, after all, be zombies.

Notes

This short article gave rise to a much longer symposium entitled "zombie earth" in a later issue of this journal.

[1] The precise difference between causal and formal processes is itself a matter of some controversy, which I need not go into here.

[2] This is of course not what the word ‘zombie’ really means, but this usage is now part of the jargon of philosophy of mind.

[3] Daniel C. Dennett refers to this as the ‘cassette theory’ (Dennett, 1978).

[4] I am indebted to Jonathan Shear for thinking of this ingenious example, and for also pointing out that, to the extent that the Terminator is supposed to be a zombie-like automaton, the device makes no sense in the movie either.

[5] This approach, which could be dubbed ‘speculative hermeneutical analysis’, is similar to the empirical hermeneutical analysis pioneered by Julian Jaynes (1976), whose reading of the _Iliad_ and other early texts suggested to him the existence of a pre-conscious ‘bicameral’ mind. In both cases, an inferential link is established between the nature of mind and kinds of possible language games observed.

[6] Like mystics, we all have a highly metaphorical phenomenological language for describing the variations in our states of consciousness. We talk about feeling ‘fogged in’ or ‘sharp’, and we understand each other. Their problem is that they must adapt the language of non-mystics to their purpose.

References

Dennett, D.C. (1978), ‘Are dreams experiences?’ in _Brainstorms_ (Montgomery, VT: Bradford Books)

Flanagan, O. (1991), _The Science of the Mind_ (Cambridge, MA: MIT Press), p. 309.

Jaynes, J. (1976), _The Origin of Consciousness in the Breakdown of the Bicameral Mind_ (Boston: Houghton Mifflin Co.).

Kirk, R. (1974), ‘Sentience and behavior’, _Mind_, 83, p. 60.

Nagel, T. (1986), _The View from Nowhere_ (Oxford: Oxford University Press).

Searle, J.R. (1992), _The Rediscovery of the Mind_ (Cambridge, MA: The MIT Press), p. 71.

JCS Home Page