Zombie Killer

by Nigel J.T. Thomas

Published in a slightly different form in:
S.R. Hameroff, A.W. Kaszniak, & A.C. Scott (Eds.): *Toward a Science of Consciousness II: The Second Tucson Discussions and Debates (*pp. 171–177). Cambridge, MA: MIT Press, 1998.

Philosopher's zombies are hypothetical beings that are behaviorally and functionally (or perhaps even physically) equivalent to (and thus indistinguishable from) human beings, but which differ from humans in not having conscious (or, at least, qualitatively conscious) mental states. They have the same information processing capacities that we humans have, and, because of this, a similar capacity to form cognitive representations and perhaps even to enter into intentional states, but they are not conscious because they do not have sensations, or *qualia* as the jargon has it. A zombie can tell you that the rose before it is red, and it will wince and hastily withdraw its hand if it touches a hot stove; however, unlike us, it never experiences the quintessential redness, the 'raw feel' of red, or the awfulness and misery of burning pain.

The zombie concept is significant because it would seem to be inconsistent with certain widely assumed doctrines about the mind. Most cognitivists assume the truth of *functionalism,* the view that mental states may all be identified with functional states, usually understood as (connectionist or symbolic) computational states, of a cognitive system. More generally, nearly everyone who hopes for a scientific account of the mind assumes the truth of *physicalism*, the view that the mind can be explained entirely within the terms of ordinary physical science, either in its present state or after some anticipated future advance. But if zombies functionally equivalent to conscious humans are a real conceptual possibility (they do not have to actually exist) then functionalism must be false, because we are admitting that two functionally indiscernible beings could be mentally different – one is conscious and the other is not. Thus there *must* be more to mentality than cognitive functioning. Likewise, if zombies *physically* equivalent to conscious humans are a possibility, then physicalism must be false. In what follows, I shall refer to those who are inclined to argue this way as "zombiphiles", and to this argument as "the zombiphile argument" (some versions are also known as the "absent qualia" argument).

Many thinkers seem to accept, gleefully or grudgingly, that zombies really are conceptually possible, and that the zombiphile argument thus goes through. However, I shall argue that, when certain implications of the zombie concept are carefully examined, zombies are revealed as either failing to support the zombiphile argument, or as simply impossible, conceptually contradictory. In what follows I shall concentrate on functionally equivalent zombies (i.e. equivalent in causal structure). Since we may safely take it that physically equivalent systems are also functionally equivalent (although not *vice-versa*), any conceptual problems with functional zombiphilia will also afflict the physical version. Physicalism will be vindicated along with functionalism.

Strict Equivalence

The concept of functional equivalence, however, is troublesome. In an important sense, it is extremely unlikely that any actual person could be entirely functionally equivalent to any other. After all, mental skills and propensities (and even brain topography) differ markedly from one person to another. Furthermore, as there is no uncontroversial principled and objective distinction to be drawn between program and data (Winograd, 1975), functional architecture and mental content, or even brain structure and the neural embodiment of particular memories, people can be said to differ functionally one from another just in virtue of the fact that they have, inevitably, had different experiences and thus have different memories and attitudes. But if the zombiphile argument is to work it would seem to be necessary to construe equivalence very strictly. If there is *any* functional difference between a zombie and a conscious human then it will always be open to the functionalist to argue that it is precisely this difference that makes the difference. If it should be discovered that there really are such beings living in our universe – and Moody (1994) argues persuasively that, under certain circumstances, characteristic and systematic behavioral differences between such zombies and conscious humans would inevitably appear – functionalists would surely be justified in treating this as an exciting *research opportunity* rather than a disaster; it would greatly facilitate the discovery of the *actual* functional basis of consciousness.

Thus Chalmers' (1996) way of setting up the zombiphile argument would seem to be the least problematic. He asks us to consider a 'zombie possible world', a parallel universe that remains at all times indiscernible from our universe in every 'external' respect, containing a physical (and functional, and behavioral) 'zombie twin' of every conscious person in this world, a being who looks, and functions, and lives exactly the same life, as each of us, but without consciousness.

In Chalmers' zombie world, some of the zombies will inevitably *claim* to be conscious. After all, my zombie twin behaves just like I do, and I make such a claim. My claim is true, but what of the zombie's? Is it false, or is there some way of interpreting it as true? Or has it no truth value at all? We shall see that any line that zombiphiles take on these questions will get them into serious trouble.

Falsity

If someone speaks falsely, they are either lying (i.e. intentionally stating things they disbelieve) or mistaken. However, since, *ex hypothesis*, my zombie twin is cognitively indiscernible from me, it cannot be lying when it claims to be conscious. Lying, after all, presupposes having an intention to lie, and if I do not have such an intention, neither does my cognitive doppelganger.

Furthermore, telling a lie surely involves cognitive mechanisms different from those involved in speaking sincerely. If, for example, the latter involves a mechanism whereby inner belief representations are converted into articulate form, lying, at a minimum, must either involve an extra or alternative mechanism whereby they are also negated, or it must apply the articulation mechanism to representations of certain *disbelieved* propositions (something I am certainly not doing when *I* claim to be conscious). In any case, my zombie twin cannot be both lying about being conscious *and* be cognitively indistinguishable from me.

But suppose the zombie genuinely but mistakenly *believes* that it is conscious. Its claims will

not be lies, and articulating them will involve exactly the same intentions and cognitive mechanisms that I employ in expressing my unmistaken belief. Here we must consider the mechanisms of belief formation. Do I and my zombie twin *infer* that we are conscious from our mutual observation of something that is reliably correlated with (or even sufficient for) consciousness in this world, but is not so correlated in the zombie world? Sometimes perhaps, but this had better not be the only way we know about our consciousness, because we could not then discover the correlation (or sufficiency). Conceivably consciousness (and the correlation) might gain its place in our conceptual repertoire as a non-observational term of some *folk theory*, but the zombiphile must surely reject this suggestion, because it leaves the door wide open to standard eliminativist moves (Churchland, 1979): i.e. to the possibility that consciousness, like phlogiston, just does not exist, that *we* might be zombies. Furthermore, given the notorious difficulty of integrating it into our scientific world view, consciousness would make an unusually appropriate target for such elimination. But if consciousness does not (or even might not) exist, if *we* might be zombies, then the zombiphile argument fails to show that functionalism might not fully explain *us*.

Thus zombiphiles normally (and plausibly) insist that we know of our own consciousness directly, non-inferentially. Even so, there must be *some* sort of cognitive process that takes me from the *fact* of my consciousness to my (true) *belief* that I am conscious. As my zombie twin is cognitively indiscernible from me, an indiscernible process, functioning in just the same way, must lead it from the fact of its non-consciousness to the equivalent *mistaken* belief. Given either consciousness *or* non-consciousness (and the same contextual circumstances: *ex hypothesis*, *ceteris* are entirely *paribus*) the process leads one to believe that one is conscious. It is like a stuck fuel gauge, that reads **FULL** whether or not there is any gas in the tank.

Such a process, like such a gauge, is worse than useless: it can be positively misleading. If the process by which we come to believe that we are conscious can be like this, we can have no grounds for confidence that we ourselves are not zombies (unlike the empty car, there will be no behavioral evidence to indicate otherwise). But (as before) if *we* might be zombies the zombiphile argument has no bite. If mistaken zombies are possible, the whole motive for ever considering such beings is undermined.

Truth

But perhaps there is a way of construing the zombie's claims as true. Although they sound like claims to be conscious like us, perhaps they are not. Perhaps zombies and humans attach different meanings to the relevant words. Moody (1994) suggests that zombies should be thought of as discussing not consciousness, but rather *consciousness*^{z}, and likewise for other mental words. However, we are now entitled to ask whether we can give any coherent interpretation to this notion of consciousness^{z}, and the zombie's use of it.

At first sight it might seem that we can. The plausibility of Moody's move surely stems from the same presupposition that makes zombies themselves seem conceivable, the widely accepted view that the aspects of mental life to do with the processing of information and the control of behavior, are conceptually (and perhaps, in principle, actually) separable from the subjective, experiential, qualitative aspects. Zombies are supposed to have the former, but not the latter. Thus, Moody tells us,

"conscious^{z}" ("conscious" as used by a zombie) will mean simply "responsive to the environment". By contrast, "conscious" as humans use it is supposed *also* to indicate the presence of the *qualia* which normally accompany such responsiveness in us.

However, the very fact that we can explain the putative difference between consciousness and consciousness^{z} carries the seeds of incoherence. If *we* can express the difference, so can zombies, and the zombie twin of anyone who explicitly claims to be conscious in the full sense, to be more than just environmentally responsive, will also make such a claim. It is not possible to construe *this* as true just by strewing more superscripts around. I can see no hope of motivating a distinction between *qualia*^{z}, for example, and qualia in the 'ordinary' sense: qualia are meant to be precisely those aspects of consciousness that are subjective and experiential and *not* informational and behavioral. But even if we *did* find a way to draw some such distinction, the problem would simply iterate: zombies would always be as ready as their human counterparts to lay claim to qualitative consciousness in its fullest, least attenuated sense, but they could not be telling the truth, or they would not be zombies.

Meaninglessness

Perhaps, however, a zombie's claims to be conscious are neither true nor false. Despite their indicative form, perhaps they have no truth value, being, in effect, without meaning. This may initially seem quite plausible because, after all, the relevant referents for words like "qualia", "consciousness" and their cognates do not exist anywhere in the zombie's universe, so the words as the zombie uses them do not refer to anything.

However, mere non-referring terms will not get us meaninglessness, at least not the right kind. It is controversial whether the fact that there are no jabberwocks makes an assertion like "I have a jabberwock in my pocket" meaningless or just plain false, but in either case it seems clear that if I *do* assert it I am either lying or mistaken. It is not like saying, "Blieble blieble blieble," or even, "I have a blieble in my pocket". After all, "jabberwock", despite having no referent, is not truly meaningless: it does have a sense of sorts – if we know our *Alice* we 'know' that jabberwocks are fearsome beasts, can be slain with vorpal swords, are fictional, etc. – and it is our grasp of this sense that allows us to reject the claim. Surely "consciousness" is similarly going to have sense for a zombie (if anything does). My zombie twin hears, reads and utters far more about consciousness than I (or it) ever do about jabberwocks, and most of this at least appears to be a good deal more intelligible than *Jabberwocky*. If a zombie can know or assert things at all, its claims to be conscious are meaningful enough to run us into the sorts of problems already discussed.

But perhaps a zombie *cannot* know or assert things at all. Perhaps *nothing* that it says, or even thinks, is meaningful. After all, Searle (1992) has argued, on quite independent grounds, that intrinsic intentionality, the meaningfulness, of our thought (upon which he takes the meaningfulness of our linguistic behavior to depend) is itself dependent upon our consciousness. If we accept this "connection principle", or something like it, (and I think there is a lot to be said for doing so), then it would seem that we not only can, but indeed *must*, acknowledge that zombie speech acts in general, and, in particular, zombie claims to be conscious, are without meaning. This would, finally, seem to provide a coherent account of the truth value of zombie claims to consciousness.

However, the zombiphile is now committed to saying that although many of the sounds and inscriptions that I make are meaningful, when my zombie twin makes exactly the same sounds or inscriptions in identical circumstances, and with exactly the same effects on its fellows, these noises and marks have no meaning whatsoever. This seems to conflict with basic notions about what it is for language to be meaningful. The zombie concept (and the epiphenomenalism which it bolsters) has a *prima facie* plausibility inasmuch as we think of consciousness as something that occurs entirely privately, completely 'inside our heads', and is thereby causally inefficacious. But whether or not this is the right way to think about consciousness, it is certainly not the right way to think about language. The meaningfulness of my utterances may depend in part on what happens inside my head, but it also depends just as crucially on what goes on in other people's heads, and what goes on between people: on both the causes and the effects of utterances. Epiphenomenalism about language should have no appeal at all. The meaningfulness of language is a social, not a purely internal, mental matter.

Consider, after all, my zombie twin in its social world, which is *ex hypothesis* just like mine. If it is greeted, it will reply politely (at least, it is as likely to do so as I am); if it is asked a question it will generally make an appropriate reply; if it is asked to do something, it will frequently either seem to comply, or else give reasons why it will not; if it is told some fact, its future behavior will (*ceteris paribus*) be appropriately adjusted. Likewise, the utterances produced by the zombie will have the same sorts of effects on its fellows. Language apparently serves just the same social functions in the zombie world as it does in ours. How, then, can we say that it is not meaningful for them? (Chalmers (1993) would seem to agree.)

More importantly, it would seem perverse to deny truth values wholesale to zombie assertions. Suppose you ask how many people are in my house, and I say, "Five". I could, for some reason, be lying, or I could be mistaken (someone is hiding), but, unless you have good reason to think that some such thing is happening, you will be wise, and likely, to assume that the information I have conveyed is accurate, that I am speaking the truth, and plan accordingly (bring five gifts, say). Surely our zombie twins are in an exactly equivalent position. My twin would lie where I would (from cognitively equivalent motives, and employing equivalent mechanisms) and would be mistaken in the same circumstances as I would, and your twin would catch it out just whenever you would catch me, and on the same basis. However, normally (and justifiably) your twin would respond as if reliable information had been conveyed: i.e as if my twin had told the truth. Can the zombiphile reasonably deny that it is right to do so? If not, the case for zombies is lost.

Consciousness, Intentionality, and Science

If zombies are impossible, it is open to us to take the perfectly plausible position that complex social and linguistic interactions are impossible without consciousness. Thus, giving up zombies need not mean giving up Searle's principle of connection between consciousness and the intrinsic intentionality that underpins linguistic meaning. Indeed, we might well go further than Searle and say that qualitative consciousness and intentionality (the *aboutness* or *directedness* of our thoughts, linguistic or otherwise) are best seen as different aspects of the same thing. "Intentionality" is essentially a term of art introduced to differentiate one key aspect of the complex pretheoretical notion of consciousness.

But consciousness certainly remains very puzzling. The flip side of zombiphile mysterianism is the quite unfounded confidence of most cognitive scientists that intrinsic intentionality need not concern them: either they think there is no such thing, or they assume there is no problem (or, at worst, a merely 'tricky', technical problem) in understanding how brain states or functional states of machines could be inherently meaningful. In fact it is hard to understand how *any* physical or functional state could be inherently *about* anything, but it also quite clear that our thoughts *are* about things. The problem of intentionality is *hard*. Nonetheless, unlike Chalmers' "hard problem" – the problem of qualia in the context of zombiphilia – it is not deliberately constructed so as to be insoluble in ordinary scientific terms. A scientific understanding of consciousness is much more likely to follow from a serious engagement with intentionality than from the fascinated contemplation of raw feels.

Acknowledgments

An earlier version was presented at *Toward a Science of Consciousness* (Tucson II) Conference, Tucson Arizona, April 12 1996. [Abstracts published in: *Consciousness Research Abstracts* (2) 1996 pp. 59–60; and *Proceedings of the Eighteenth Annual Conference of the Cognitive Science* Society (Erlbaum, 1996).]

This version largely written whilst a participant in the 1996 NEH Summer Seminar: "Metaphysics of Mind". Thanks for comments and discussion to: Dave Chalmers; Jessica Wilson; the seminar director, John Heil; and the seminar participants, especially Eric Saidel, David Pitt, Jim Garson, Tony Dardis, and Steve Schwartz.

References

Chalmers, D.J. (1993). "Self-Ascription Without Qualia: a Case Study." *Behavioral and Brain Sciences* (16) 35–36.

Chalmers, D.J. (1996). *The Conscious Mind.* Oxford University Press.

Churchland, P.M. (1979). *Scientific Realism and the Plasticity of Mind.* Cambridge University Press.

Moody, T.C. (1994). "Conversations with Zombies." *Journal of Consciousness Studies* (1) 196–200.

Searle, J.R. (1992). *The Rediscovery of the Mind.* MIT Press.

Winograd, T. (1975). "Frame Representations and the Declarative/Procedural Controversy." In: *Representation and Understanding.* D.G. Bobrow & A. Collins (Eds.). Academic Press.