ZOMBIES, EPIPHENOMENALISM, AND PHYSICALIST THEORIES OF CONSCIOUSNESS

Andrew Bailey

Department of Philosophy

The University of Guelph

Guelph, Ontario N1G 2W1

Canada

(519) 824-4120 x53227

abailey@uoguelph.ca

Abstract

Philosophical zombies are theoretically stipulated creatures which are outwardly, behaviourally, even physiologically indistinguishable from normal human beings, but which lack consciousness. The possibility of zombies is appealed to in contemporary philosophy of mind to show either a) that consciousness is not essential for intelligence, or b) that physicalism must be false and some form of metaphysical dualism or neutral monism true. In this paper I argue that the notion of zombies which are *physiologically* identical with human beings while lacking consciousness is incoherent, and so physicalism with respect to consciousness should be considered immune to zombie-based attacks (and is in fact bolstered by their failure). However, I argue that the notion of functional zombies is *not* incoherent in the same way and that these two results favour a biologically, rather than computationally, oriented approach to the problem of consciousness.

ZOMBIES, EPIPHENOMENALISM AND PHYSICALIST THEORIES OF CONSCIOUSNESS

In its recent history, the philosophy of mind has come to resemble an entry into the genre of Hammer horror or pulpy science fiction. These days it is unusual to encounter a major philosophical work on the mind that is not populated with bats, homunculi, swamp-creatures, cruelly imprisoned genius scientists, aliens, cyborgs, other-worldly twins, self-aware computer programs, Frankenstein-monster–like 'Blockheads,' or zombies. The purpose of this paper is to review the role in the philosophy of mind of one of these fantastic thought-experiments—the zombie—and to reassess the implications of zombie arguments, which I will suggest have been widely misinterpreted. I shall argue that zombies, far from being the enemy of materialism, are its friend; and furthermore that zombies militate against the computational model of consciousness and in favour of more biologically-rooted conceptions, and hence that zombie-considerations support a more *reductive* kind of physicalism about consciousness than has been in vogue in recent years.

1. PHILOSOPHICAL ZOMBIES

Zombies—of the philosophical rather than the Haitian or Hollywood variety—are theoretically constructed creatures stipulated to be identical in certain respects with ordinary human beings, but lacking in other respects.[1] Perhaps the most

familiar member of the zombie family is that non-actual but (putatively) possible creature which is functionally identical with a 'normal human being' but entirely lacking in phenomenal states, in states of experiential consciousness. Such a zombie, as is characteristic of the breed, would be entirely indiscernible from the regular folks among whom it walks—as Owen Flanagan and Thomas Polger have put it (1995), it could fool even the sharpest 'mental detector'—but there would be 'no lights on' inside its head; all would be dark inside. A zombie will sometimes behave *exactly* as if it were in pain, or in love, or enjoying a movie, but could actually occupy none of those states: it could never, *ex hypothesi*, actually *feel* pain or *experience* enjoyment.

As is sometimes—but not always—noticed, the notion of a zombie is not a unitary one, and subtle variations in the construction of one's zombie thought-experiments can have important ramifications for their philosophical consequences.[2] There are at least three scales along which zombies can vary. First, there can be different ways in which zombies may be stipulated to be *identical* with normal persons: in particular, they may be quark-for-quark *physically* identically with people; they may be *functionally or computationally* identical with people; or they may merely be *behaviourally* identical or indiscernible. Second, zombies may be constructed such that they *differ* from normal people in various ways: for example, they might be postulated to lack any psychological states at all, or to be missing some subset of the psychological such as intentional or phenomenal states. Finally, there can be important differences

in the *modality* of zombie claims: to assert that creatures indiscernible from humans but lacking mentality are possible is not yet to specify whether zombies are merely logically possible, metaphysically possible, nomically or naturally possible, or physically possible. Zombies might be logically possible but nomically impossible, for example, if their existence is incompatible with laws holding in the actual world (and in nomically similar worlds) but consistent with different laws of nature that hold in other possible worlds. Similarly, zombies might be physically possible but nomically impossible if the laws of nature in the actual world outrun the physical laws—that is, if there are natural laws which are not also physical laws—and if zombies are consistent with the holding of all the physical laws but not with the holding of all the natural laws.

What, then, are zombies for—what is the point of constructing these fictional characters? Zombie thought-experiments, properly constructed, are ideal limit cases which can be used to reveal and assess the consequences of various different philosophical theories. Philosophical zombies were first introduced, by Robert Kirk in 1974, as a test case for physicalism, and more recently they have seen steady use in critiques (especially dualist critiques) of the various metaphysical theories of consciousness—behaviourism, identity theory, functionalism, non-reductive physicalism—and of various related notions, such as the coherence of the 'Strong' Artificial Intelligence project, and of biological or evolutionary theories of the mind.

In contemporary philosophy of mind there are at least two roles—not always

clearly distinguished from each other—that zombie thought-experiments are intended to play. The first, exemplified in the work of Ned Block, Owen Flanagan, Thomas Polger and Todd Moody, is the defence of what is sometimes called "conscious inessentialism" (e.g. Flanagan 1991, 309). This is the claim, roughly, that consciousness is not necessary for intelligence: two creatures could be equally well fitted for a particular environment, equally well able to form internal representations of that environment and act on them in an intelligent manner, and yet one might be conscious and the other not. The pressing question then becomes, why are *we* conscious? Why is our intelligent behaviour mediated by consciousness, when it might not have been? I shall not dwell on this topic in this paper, but one important thing to notice about this use of zombies is that (as Polger (2000) has emphasised) *these* zombie thought-experiments do not entail that consciousness must be epiphenomenal. From the possibility of creatures who are (at a reasonably high level of abstraction) functionally identical with us but lacking consciousness, it does not follow that *our* consciousness must be epiphenomenal: the occupants of the relevant functional roles in us might be conscious states but in zombies unconscious ones, and the occupants of functional roles—though *functionally* neutral (with respect to those roles)—clearly need not be epiphenomenal.

The second main role for zombies is as foot soldiers in the so-called "qualia wars." For any given, materialist theory of phenomenal consciousness, there is a zombie thought-experiment lurking which attempts to call that theory into doubt.

Behaviourism, for example, is challenged by the possibility of behavioural zombies: creatures whose behaviour is *ex hypothesi* indistinguishable from that of a paradigm normal person, but which entirely lack consciousness. Similarly, functional zombies lie in wait for functionalist theories of consciousness, and physiological zombies for central state materialists.[3] The general argument pattern can be laid out as follows:

1) If theory X were true, conscious states would be type-identical with, or at least logically supervenient on,[4] X-ish states. For example, behaviourism is the claim that mental processes—insofar as they exist at all—*are* (are identical with/are ontologically nothing over and above) behavioural dispositions; functionalism is the claim that mental states or processes *are* functional states or processes; and so on. From this it would follow that—if we hold constant the relevant background facts—it is *logically impossible* for X-states to occur in the absence of conscious states.
2) Creatures possessing X-states but lacking consciousness—X-zombies—are conceivable, and whatever is conceivable is logically possible.
3) Thus it is logically possible for X-states to come apart from conscious states, and so theory X is refuted.

If functionalism, for example, is a true theory of consciousness then functional zombies are incoherent—they are logically impossible. But, the argument goes, functional zombies are logically possible (and this fact can be established for

reasons independent of simply asserting the falsity of functionalism—e.g. through conceivability considerations). Hence, we are to conclude, functionalism is not true with respect to consciousness.

There are various ways in which one might consider critiquing such an argument, but at least for the purposes of this paper I shall assume that this general argument strategy is cogent: that is, I shall assume, if X-zombies are appropriately conceivable, then X-zombies are logically possible, and the logical possibility of X-zombies is sufficient to refute Xism as a theory of consciousness. Two important, related difficulties for arguments of this sort are a) the murkiness of the relation between conceivability and logical possibility, and b) the relationship between logical possibility and metaphysical possibility. In particular, some proponents of, for example, functionalism are happy to concede the *conceivability* of functional zombies but deny that this reveals that functional zombies are a *metaphysical* possibility (just as the 'epistemic possibility' of water not being H_2O fails to show that "water = H_2O" is not metaphysically necessary).[5] These issues are rich and complex, and this is not the place for even a cursory survey of their ramifications; although Kripke-style examples seem initially persuasive, various principled defences of these contested connections have been mounted.[6] For present purposes, suffice it to say that it is, at least, far from clear on a detailed examination of these issues that zombie arguments of the type sketched above *cannot* work; and that if they can operate as advertised then, I will argue, this will have the consequences I describe in the rest of this paper.

(Conversely, if conceivability and the appropriate sort of possibility cannot be tied together somehow, then zombie arguments—as well, presumably, as many other appeals to thought-experiments to establish metaphysical conclusions—simply become philosophically irrelevant.)

2. The Zombie Assault on Physicalism

Assuming that appeals to zombies are ever worth considering, there is one version of the zombie argument that is of particular metaphysical significance since it has implications that go beyond the truth or falsity of some theory of phenomenal consciousness. If *physiological* zombies are logically possible then this ought to be sufficient to establish not merely that some particular physicalist theory of consciousness is false but that physicalism itself is untrue in the actual world.[7] If physicalism is false with respect to consciousness then it is false *simpliciter*. Conversely, if physiological zombies are *not* logically possible, then it seems that physicalism—with respect to the mental—*must* be true.[8]

The central idea here is that, whatever else the doctrine of physicalism is, it essentially involves the claim that the physical facts, in some sense, exhaust all the facts: once God (metaphorically speaking) created all the physical facts there was nothing that remained to be done to fix the rest of the facts. This is not, or need not be, to say that all facts—economic, aesthetic, sociological and the rest—are *reducible* to facts recognizable in the language of physics; our minimal commitment is simply that all facts whatever are *fixed* once we have fixed the

facts of physics. Hence, physicalism is true just in case every fact (state, process, property, etc.) globally supervenes by logical necessity on the physical facts; any two possible worlds that are indiscernible physically and in which physicalism is true are indiscernible *simpliciter*.[9] So the logical possibility of physiological zombies—creatures physically just like us, embedded in a world physically identical to the actual world, but of whom the facts about consciousness differ—would be sufficient to refute physicalism.

Of course this implication, if it could be established, is also a result in the philosophy of mind. My physiological zombie twin is, *ex hypothesi*, cell-for-cell, microtubule-for-microtubule, even molecule-for-molecule, identical with me, but (even though the laws of physics and all the relevant physical initial conditions are also held constant[10]) is entirely lacking in consciousness; *my* consciousness, therefore, if I do have a logically possible zombie twin, must be something over and above my physical constitution.

Everything, then, hangs on whether physiological zombies are logically possible. And I am presently granting that, at least in some cases, the logical possibility of some state of affairs can be established through its conceivability. The question, then, is whether physiological zombies are, in the appropriate way, in fact conceivable. I shall argue here that they are not. Furthermore, I shall argue that the notion of a physiological zombie is actually *incoherent* and thus that zombie considerations, far from refuting it, go some distance towards *affirming* physicalism with respect to consciousness.

The reason that physiological zombies are inconceivable, in short, is that they go hand-in-hand with the doctrine of the metaphysical epiphenomenalism of consciousness in the actual world; and this supposition—the radical epiphenomenality of consciousness—is incoherent.

3. Zombies, Physicalism and Epiphenomenalism

To a first approximation, the logical possibility of physiological zombies implies the epiphenomenalism of phenomenal consciousness in the actual world. The supposing of physiological zombies is the supposing (conceiving, imagining) of a possible world specified by the subtraction of all consciousness from our world, or at least from a part of it, while holding *everything* physical constant.[11] Since we hold everything physical constant, then—by stipulation—every physical event will happen in the zombie world just as it does here. This will include, among other things, all the neural events occurring in my zombie twin, all the information-processing taking place within my zombie twin, all my zombie twin's external behaviours (including his linguistic behaviours), and so on. Thus, on the face of it, consciousness is not required for everything to happen in the actual world just as it does, and so consciousness is radically epiphenomenal (in the actual world).

There are two reasons to moderate this conclusion however. The first is that it may be the case in the actual world, consistently with the logical possibility of physiological zombies, that certain events are causally *overdetermined* such that they have both a sufficient physical cause and an additional non-physical cause. If

that were so, then consciousness would not be required for everything to happen in the actual world just as it does but nevertheless phenomenal consciousness would not be causally inert—it would have causal powers that map on to existing sufficient physical causes.[12]

Considerations to do with overdetermination leave the fundamental issue unchanged, however. Consciousness may not be epiphenomenal, on this scenario, but it is still what we might call 'quepiphenomenal': it is causally *irrelevant* in that its presence or absence makes no difference to how everything (non-phenomenal) goes in the world. And it is this irrelevance, rather than the full-blooded thesis of epiphenomenalism, that does the work in the arguments to follow: the problems arise because of the implausibility of supposing that the elimination of consciousness would make no difference at all to how things go in the physical world.[13] So, so far, the logical possibility of physiological zombies entails that consciousness is either epiphenomenal or quepiphenomenal in the actual world; backing away from my barbarous neologism I shall phrase my arguments below primarily in terms of epiphenomenalism, but closely analogous versions go through for quepiphenomenalism.[14]

The second consideration affecting the entailment of conscious epiphenomenalism in the actual world by the logical possibility of physiological zombies is more subtle. It is generally agreed that physicalism is a contingent doctrine: physicalists typically hold not only that physicalism *might* have turned out to be false had the world been different than it actually is, but even that it

might *still* be discovered to be false (though the physicalist bet is that it will not). The zombist a priori argument against physicalism thus attempts to show that physicalism *cannot* be true—that it is incoherent even to suppose it true of the actual world—since any weaker a priori conclusion can be accommodated by the physicalist.[15]

What must physiological zombies be for this anti-materialist argument to work? They must be such that they are physically just as the physicalist supposes *we* are, but lacking consciousness. (After all, that possible worlds with quite different physical configurations or laws than our own might contain 'zombies' is of no relevance to the truth of physicalism in the actual world.) But they need not be physically just the way we *actually* are, since physicalism may be false of the actual world. When Chalmers and others write of zombies being physically *identical with* or *indiscernible from* actual inhabitants of the actual world they are writing a little loosely; what is required is merely that they be fixed as the way physicalists *suppose* we are. If, on the physicalists' best and most optimistic assumptions about the actual world, fixing the physical does not fix the phenomenal—i.e. if physiological zombies are logically possible even given a physicalist account of the physical facts—then physicalism cannot be true. But once we have drawn this conclusion, it seems to be open to the anti-materialist to reject some of the physicalists' assumptions about the physics of the actual world.

In particular, the anti-materialist might speculate that physics is in fact not causally closed—that not every event has a sufficient physical cause[16] and that the

removal of consciousness from the actual world would therefore *change* the way things go. For example, interactionist dualism may be true. More subtly, it could be that what Chalmers calls panprotopsychism is the case, whereby the intrinsic nature of the physical is composed of phenomenal or proto-phenomenal properties, and physical laws derive from more basic laws connecting the intrinsic properties. On this scenario, it would not be possible to remove the phenomenal while leaving the physical just as it is, as the physical is in some sense *constituted* by the phenomenal.[17] In short, if physicalism is false consciousness might not be epiphenomenal or even quepiphenomenal.

Zombie arguments against physicalism thus presuppose a thesis which is only contingently true, the causal closure of the physical; and if it is false, consciousness need not be epiphenomenal even if it is non-physical. But the making of this assumption *is not itself contingent* for anti-materialists utilizing zombie arguments. Physiological zombie arguments cannot even be formulated without this assumption, since without the causal closure of the physical, phenomenal consciousness could not be subtracted from the world while leaving everything else the same—the removal of consciousness would leave causal gaps that would change the sequence of physical events (perhaps eliminating pain qualia would alter human pain behaviour, for example).[18] The dialectic thus runs as follows:

a) Suppose physicalists are right about the causal closure of physics. Then—argues the zombist—physiological zombies are logically possible;

hence physicalism must be false and consciousness epiphenomenal.

b) Suppose physicalists are not right about causal closure. Then physiological zombies are *not* possible—it is not logically possible to remove consciousness from the actual world and leave the physical history the same—so zombie attacks on physicalism cannot get off the ground. But that doesn't matter, for the anti-materialist, since to establish the failure of the causal closure of physics with respect to the phenomenal is *already* to establish the falsity of physicalism.

What the anti-materialist cannot do, however, is assert that physicalism is false *because* of the logical possibility of zombies *and* hold that consciousness is not epiphenomenal. The conceivability of physiological zombies *entails* the epiphenomenalism of consciousness (even though the falsity of physicalism does not).[19]

What about the following, though? Suppose for the sake of argument the actual world isn't causally closed, and suppose also that the world nevertheless *might have been* causally closed under physics, even though it isn't. So there is a possible world, *W*, physically rather like this one, and containing the distribution of phenomenal consciousness with which we are familiar in this world, with the important difference that *W* is causally closed under physics. Now, from the perspective of *W*, as it were, physiological zombies are possible: that is, George Bush_w might have a zombie twin, even if the real Bush could not.[20] Thus, in *this* sense, the possibility of physiological zombies is consistent with the non-holding

of conscious epiphenomenalism in the actual world. But epiphenomenalism still holds in *W*. So the result still obtains: in any world 'with respect to which' physiological zombies are conceivable, consciousness in that world is epiphenomenal. If this is such a world, then consciousness is epiphenomenal in this world.[21]

There is one final retort to consider before moving on: "an interactionist dualist can accept the possibility of zombies, by accepting the possibility of physically identical worlds in which physical causal gaps go unfilled, or are filled by something other than mental processes" (Chalmers 2004, 184). Take a world that is not closed under physics, where some of the causal work is done by phenomenal states; remove the phenomenal, but then make some further stipulation about the world to ensure the causal gaps that would otherwise remain are somehow bridged and so that the physical can continue along its way indiscernible from the target world. This seems at first sight sensible. But consider: Suppose we decide to fill the causal gaps—what are we to fill them with? The zombie world is to be physically indiscernible from its target world, so we cannot change the physics; and it is to lack states of phenomenal consciousness, so we cannot use those. If we cannot fill the gaps with either matter or spirit, then what candidate substance remains? Suppose we elect not to fill the causal gaps but simply stipulate that the physical events continue to occur as they do in the target world. This cannot be done without changing the physics. As Chalmers himself is often at pains to point out, the characterization of the physical is

structural and relational; what makes an electron an electron, as far as the physical sciences are concerned, is the way it is embedded in a set of law-like causal relationships with other entities. Chalmers says that "there is nothing metaphysically impossible about unexplained physical events" (2004, 184), which is indeed the case, but it *is* impossible for two physical events, one connected by natural laws to other event-types and the other not so connected and hence 'unexplained,' to be *the same* physical event (i.e. members of the same physical event-type).

Suppose it turns out—as I shall argue in a moment—that physiological zombies are *not* conceivable: is this consistent with the contingency of the doctrine of physicalism? It is. First, we have seen that the impossibility of physiological zombies might be due either to the truth of physicalism *or* the failure of causal closure; so defeating the zombist anti-materialist does not rule out, say, interactionist dualism. Second, even if causal closure is true but zombies impossible, the universe might contain as-yet undiscovered—or even undiscoverable—epiphenomenal non-physical phenomena, such as ghost particles or disembodied spirits.[22] The fundamental reason that physicalism is an empirical doctrine is that it cannot be established wholesale but only piecemeal, by showing that one phenomenon after another is physical; this is a process that might continue indefinitely, yet only when it is completed could we be sure that physicalism is true.

This type of contingency, however, does not call into doubt the progress the

physicalist has made so far: we know that water is, beyond a shadow of a doubt, an entirely physical substance, that electro-magnetism is a physical force, and that geological and biological processes are nothing over and above their physical substrate (albeit that that relationship is a complex one of supervenience, and not one of identity). To show that human phenomenal consciousness is also physical (i.e. fixed by the physical facts alone) would be an important move forward in this progress; and, given causal closure, it is this which the inconceivability of physiological zombies would establish.

4. The Falsity of Phenomenal Epiphenomenalism

So physiological zombies are conceivable only if phenomenal consciousness is metaphysically epiphenomenal. It is not enough merely to conceive of a possible world, physically similar to this one, in which consciousness is epiphenomenal: the very logical possibility of physiological zombies (with respect to the actual world) requires, since it entails, that phenomenal consciousness really *is* epiphenomenal in the *actual*, non-zombie world. But this, I shall argue, is false—phenomenal consciousness is not, as a matter of fact, epiphenomenal. That is, in order to show that physiological zombies are inconceivable, I need not establish that epiphenomenalism is itself inconceivable (i.e. logically impossible)—just that it is certainly false of the actual world.

So why do I say that phenomenal epiphenomenalism is false? After all, it perhaps seems on the face of it as if it could be true—as if it is not obviously in

tension with the known facts, however peculiar it would be if it turned out to be actually the case.[23] But on reflection there are at least three phenomena with which the epiphenomenalism of consciousness cannot be reconciled: first-person reporting of consciousness, the semantics of terms for consciousness, and the emergence of consciousness in the history of life on earth. Arguments resembling those I shall appeal to in this section have, to a greater or lesser extent, occurred to others before (and I give references throughout to the closest examples I am aware of), but I hope here to bring them together, distil them into their strongest form, and place them in the wider context of the implications of zombie arguments for the philosophy of mind.[24]

The first of these three difficulties for epiphenomenalism might be called the Reporting Problem. If consciousness is epiphenomenal then it has no effects; in particular, it has no effects on those organisms whose consciousness it is. Thus, if physiological zombies are conceivable, it follows that my own 'reports' of my conscious experience—my complaints that I am in pain, or my assertion that I am currently viewing a red field, for example—are not caused by that experience at all. Indeed, not only are they not caused by my experience but they are caused by something *entirely different*, something physical. Furthermore, the causal chain which brings these reports about is one which only contingently 'tracks' the experience that they purport to be about—physiological zombies, after all, make precisely the same reports as normal human beings do, in precisely the same physical circumstances, but lack consciousness.

One version of this problem is particularly amusing. If phenomenal epiphenomenalism is true then if philosophers are baffled about the phenomenon of consciousness it cannot be *because* they have noticed that consciousness is a mystery.[25] The mental state of bafflement, like all mental states in the actual world, is instantiated by the brain; and the brains of physiological zombies are neuron-for-neuron identical with those of regular people. Thus, my zombie twin will have all the psychological states that I would—though, certainly, the *phenomenal character* of those states will be different (i.e. absent) for it—and so my zombie twin will report being *just as baffled* by the mystery of consciousness as I am. A zombie Mary, leaving her black and white room and encountering colour for the first time, will make *exactly the same* expressions of delight and surprise as regular Mary.[26] A zombie Chalmers will be just as adamant that its own 'phenomenal consciousness' is over and above the physics of its brain, and a zombie Dennett will be exactly as certain that zombies are preposterous. Finally, since the philosophical bafflement of zombies is not caused by consciousness, and since zombies and regular people are causally identical *ex hypothesi*, *my* philosophical bafflement 'about consciousness' is not caused by consciousness either (if epiphenomenalism is true).

So, if phenomenal epiphenomenalism were true, the verbal and other behaviours, both internal and external, that we took to be reports of consciousness would turn out just not to have the characteristics of genuine reporting with respect to consciousness; worse, they *would* stand in an

appropriate relation to be genuine reports of something *other* than consciousness (certain species of neural state with which they are defeasibly correlated in a counterfactual-supporting way).[27] But of course none of this is right: when I report that I am baffled about consciousness it is *consciousness* that I am baffled by; when I complain of a nagging itch, it is the *itchiness* that nags. We *do* sometimes make genuine first-person reports about phenomenal consciousness (though of course in some sense 'we' *might* not have done); phenomenal epiphenomenalism is inconsistent with this fact; so phenomenal consciousness cannot be epiphenomenal. If phenomenal consciousness cannot be epiphenomenal, then physiological zombies are inconceivable.

It bears emphasising that the Reporting Problem is distinct from what David Chalmers calls the Paradox of Phenomenal Judgement. The Paradox of Phenomenal Judgement amounts to the claim that we cannot know about our own conscious states since our beliefs (judgements) about those states do not stand in the appropriate causal or explanatory relation to them. Chalmers attempts to defang this paradox by arguing that our knowledge of experience is properly seen, not as grounded in any causal relation, but in a "more immediate relation" (1996, 198). However, the Reporting Problem is not a problem about self-knowledge—which may or may not be mediated causally—but is a problem about reporting, which surely does involve appeals to causal-historical chains. Put it this way: even if my report is caused by some introspective belief that a zombie cannot share (perhaps since the content of that belief is partially

constituted by phenomenal experience, as Chalmers argues (2003)), the phenomenal aspect of that belief is irrelevant to the reporting—is not what is being reported—since exactly the same reports would be issued even if the phenomenal aspect were absent. Analogously, if I report that I once climbed Mount Kilimanjaro I am not thereby reporting that I climbed it wearing a kilt, since I would have made the original report whether or not the latter aspect of the event were the case.

The second problem for phenomenal epiphenomenalism is what might be called the Semantic Problem. If consciousness is epiphenomenal then it has no effects; in particular, it has no causal effects on language behaviour. Clearly, then, if consciousness is epiphenomenal, standard causal-historical, information-theoretic, counterfactual or teleosemantic accounts of intentionality will not work for phenomenal terms: for example, we cannot refer to conscious states in virtue of their being the 'normal' or 'proper' cause of our detection of them, since epiphenomena are not any kind of causes at all (let alone 'proper' ones). Furthermore, interpretationist semantics of the Davidsonian or Dennettian type will also apparently fail to give us a proper semantics for phenomenal terms, since zombies will satisfy exactly the same set of interpretations as normal human beings despite lacking any phenomenal states.[28]

Chalmers (1993, 2003), among others, has tried to argue that reference to conscious states (which, like self-knowledge, must be distinguished from reports of the presence of these states) might be mediated by some sort of non-causal

first-personal *acquaintance* with them: but from this it follows, as Chalmers recognises, that zombies must mean something different when they talk about mental experience than we do—probably, in fact, it must be that when zombies use phenomenal language they mean nothing at all.[29] Apart from well-known 'private language' type difficulties with this kind of story about reference, an acquaintance-based semantics would have the puzzling consequence that, not only are qualia epiphenomenally supervenient, but a certain class of *meanings* fail to supervene at all: that is, on this view, fixing all the physical, social and cognitive facts about linguistic behaviour will, sometimes, fail to fix or even to constrain the meanings of those utterances (even though there is nevertheless a fact of the matter about what they mean). Two individuals—a zombie and a regular person—can be members of functionally identical language communities, have just the same (non-phenomenal) cognitive architecture, be identical with respect to their environmental and historical circumstances, can engage in overtly identical speech acts with completely similar communicative success, and yet, for the epiphenomenalist, they might mean utterly different things by their utterances. Not only will certain mental facts fail to supervene on the physical, if physiological zombies are conceivable, but many *linguistic* facts will not supervene either. Even many normally clear-sighted dualists, such as David Chalmers, do not seem to have faced fully up to this dubious implication of phenomenal epiphenomenalism.

The third problem for phenomenal epiphenomenalism is the Emergence

Problem. If consciousness is epiphenomenal then it has no effects; in particular, it has no effects on an organism's evolutionary fitness.[30] It follows directly that consciousness cannot have been selected for, and this raises the puzzling question of *why*, if consciousness were epiphenomenal, any creature would be conscious. (It is tempting to suppose that the presence of consciousness makes pain just that bit more pressing, or lust just that bit more imperative, but a moment's thought shows that physiological zombies are, by stipulation, exactly as ready to respond to pain stimuli or a receptive mate as we are.) The mystery of consciousness is thus compounded. Not only must we ask *how* are we conscious, but we are now faced with the question: *Why* are we conscious? Furthermore, there seems on the face of it to be no reasonable hope of answering this latter question: the most natural way of responding to any Why question is with a functional-teleological or mechanistic explanation, but any such response is already ruled out by the assumption of epiphenomenalism. It's not clear even what would *count* as a satisfactory answer to the why question when it is posed in this way (as opposed to the way consciousness inessentialists such as Owen Flanagan want to put it).

Consider, for example, the proposal that consciousness, though evolutionarily inert, is to be treated as a kind of spandrel of the brain—that is, the view that consciousness, though not itself a contribution to fitness, is an essential companion to attributes (presumably neural attributes) which *do* confer an evolutionary advantage. In such a case, there *would* be an evolutionary

explanation for consciousness, in a way analogous to the way that we can explain the presence of the triangular spaces—spandrels—formed between arches which meet at the centre of a dome roof entirely in terms of the architectural benefits conferred by such arches. However, such evolutionary explanations are unavailable to those who accept the logical possibility of physiological zombies, since this possibility is exactly the claim that consciousness is *not* a (physically, or *a fortiori* logically) necessary accompaniment of anything that confers evolutionary fitness: an organism would be exactly as fit without consciousness, whereas a dome could not—as a matter of geometrical necessity—be as well supported by arches if it lacked spandrels.

The Emergence Problem, however, is not necessarily tied to evolutionary considerations. In its purest form it is this: Suppose that no non-living thing, in the actual world up to this date, is conscious; suppose that single-celled organisms and other very simple and old forms of life are not conscious. Then at some point in the history of life *something must have happened* to bring about the emergence of consciousness. But everything that happened physically in the actual world up to the present moment, given the causal closure of the physical to which the zombist is committed, happened for sufficient physical reasons (or perhaps as a result of quantum randomness): in that sense, there is some explanation for it—some naturalistic, even if probabilistic, *reason* why it happened. But, if consciousness is epiphenomenal, it must uniquely fall outside this web of otherwise universal naturalistic explanation. To what sort of

explanation might its emergence be susceptible then? Why is its occurrence correlated, apparently, with a certain sort of neural complexity? Merely appealing to extra-physical 'psychophysical laws of nature'[31] will not help—why *these* laws, we might ask, and not others? Apart from, say, appeal to some sort of divine agency, no further explanation seems possible.

This is not, notice, merely an epistemic problem. The situation is not merely that we cannot (yet) provide an explanation for the historical emergence of consciousness; the problem is that that emergence itself is *inexplicable*—if consciousness is epiphenomenal then there *is no reason* for the existence of consciousness, not even a merely contingent historical reason. This is a claim about the phenomenon of conscious itself, not merely our knowledge of that phenomenon. And it is a claim that seems most likely false of the actual world.

I have argued thus far that we have strong reasons to think phenomenal epiphenomenalism false of the actual world, and hence that physiological zombies are in fact inconceivable since the supposition of a (merely logically possible) zombie world necessarily involves the supposition that consciousness *is* actually epiphenomenal. In schematic form, (□ P □ Q), ~Q, so ~□ P; epistemically, once we come to believe that ~Q we are no longer able both to fully understand P and to conceive of a possible world in which P is the case.

5. The Unimaginability of Physiological Zombies

Presumably there is a relevant difference between merely saying the words "a

creature exactly like me except that it lacks consciousness" and actually *conceiving* of such a creature. As John Perry puts it, "[t]o show that there is a possible world meeting certain conditions, one must imagine or describe it in enough detail to be sure it is possible and meets the conditions in question" (2001, 80). Though Chalmers insists that the burden of proof is on the anti-zombist to show why the notion of physiological zombies is incoherent—a burden I have tried to discharge in the previous section—he does suggest that the *reason* the burden of proof swings this way is that "I have a clear picture of what I am conceiving when I conceive of a zombie" (1996, 99) … "the logical possibility of zombies seems … obvious to me" (1996, 97).

There are reasons to think, however, that the positive conceivability of physiological zombies is not quite as obvious an affair as it might at first seem. That is, I will now argue, not only is it incoherent to suppose (~Q and □P) together, but it may not even be coherent to suppose □P alone. The considerations supporting this latter claim are, I think, more suggestive than conclusive, but they are still worth raising.[32]

What would it be, then, to have a 'clear picture' of a physiological zombie—what kind of mental task are we being asked to perform when we are asked to conceive of a zombie? We cannot imagine a zombie 'from the inside' as it were—we cannot imagine what it would be like to be a zombie, since there *is* nothing it is like to be a zombie.[33] Zombies, therefore, must be imagined 'from the outside': but from this perspective, there is no difference at all between imagining a

physiological zombie and imagining a normal human being, no matter how much behavioural, cognitive or physical detail you conjure up. There is therefore, it seems, in principle *no way to tell the difference* between successfully imagining a zombie and failing to do so. And this is surely a problem for the zombist: for any task, including 'conceiving' tasks, the command to perform the task is simply ill-formed if there is *in principle* no difference between performing the task and not doing so.

One might respond to this by saying something like the following: I will think about my own introspective experience and suppose that a zombie is something just like me but lacking *this*. But this strategy will work only if we commit to the far from uncontroversial view that phenomenal states are partially *constitutive* of some mental content (the view that, e.g., the belief that I am currently in pain has its content at least partly in virtue of the felt pain itself).[34] If, as many suppose, the cognitive and the phenomenal are in principle separable then the 'introspective strategy' is problematic: just like us, zombies sincerely judge that they are conscious—by stipulation, they have just the same cognitive (non-phenomenal) mental states as we do, so whatever we believe or judge[35] about consciousness they will believe as well. (Zombies could never discover that they are in fact zombies, for then they would be functionally different from us and thus no longer zombies.) My zombie-twin, therefore, will be saying to himself, "I conceive of zombies as being something like me but lacking *this*," meaning its own phenomenal consciousness, but in its case the *this* will fail to refer.

Nevertheless, there is no cognitive (as opposed to phenomenological) difference between me and zombie-Andrew, so zombie-Andrew will be just as convinced of its ability to conceive of zombies as I am.

The problem now is: how do I know that I am not the zombie? The fact that I am *sure* I am not, that I feel *certain* that I undergo conscious experience, that I think the notion I might be a zombie a *ridiculous* one—none of these things make any difference, for zombie-Andrew has exactly the same reactions.[36] It's true that there *is* a difference between me and zombie-Andrew—that it feels like something to be me and like nothing to be zombie-Andrew; the problem is that this difference makes no *cognitive* difference. The strategy for imagining zombies by subtracting one's own consciousness thus undercuts itself: if it is successful, then zombies may be logically possible; but if zombies are possible then *I* would be unable (cognitively) to be sure that I am not a zombie; but zombies cannot use the above technique to successfully imagine zombies; so since I can't tell the difference between being a zombie and not, I still can't tell whether I can successfully conceive of zombies.[37]

One of the most careful defences of the imaginability of zombies comes from Robert Kirk's original introduction of the notion.[38] Kirk (1974a) describes a man—whom he calls Dan—who undergoes the progressive loss of his consciousness. First Dan loses his pain sensations: he responds to pain stimuli perfectly normally (including making the usual verbal responses to pain) but in addition he *also* expresses astonishment at the disappearance of his pain

sensations. Then, gradually, he loses all his other modes of sensation, each time remaining behaviourally, functionally, and physiologically exactly the same, and each time expressing anguished dismay at the erosion of his consciousness; eventually, all the sensation is gone, and Dan has become zombified. Kirk's argument is that each of these intermediate stages on the way to zombification is perfectly conceivable, and furthermore that in such a case the *best explanation* for what was going on would be zombification: we would be entitled in these intermediate cases to attribute partial zombiehood to Dan, since only this would explain both, say, the pain behaviour (caused by neural states) and the expressions of astonishment (caused by changes in conscious experience). Since, according to Kirk, it is plausible that we can imagine Dan as a part-zombie, there should be no barrier to taking the final small step and conceiving of him as eventually becoming all-zombie.

The problem with Kirk's argument, however, is that its premises are incoherent. First, Kirk requires, for the defence of physiological zombie conceivability, intermediate stages where Dan is still physically just the same as he always was, but nevertheless reporting the loss of his sensations: but this is impossible, given the parameters of the zombie thought experiment. Expressions of dismay and astonishment are physical behaviours, as are the relevant physical differences in the brain associated with them, and hence these are ruled out *ex hypothesi*. Second, Kirk's argument that zombification is the best explanation of what is happening to Dan presumes that the loss of sensation might *causally*

explain Dan's expressions of astonishment: but of course, this assumption is inconsistent with the conclusion which it is supposed to establish, since if zombies are possible consciousness is epiphenomenal, and if consciousness is epiphenomenal its presence or absence makes no difference to—is explanatorily irrelevant to—behaviour.

The burden of this section is to suggest that, contrary to an apparently widely shared impression, we have no firm, intuitive grasp on what it is to conceive of a physiological zombie (beyond merely the bare assertion of the propositions involved). Unlike the case of, say, mile-high unicycles or thousand-sided polygons, there seem to be no further details we can adduce to flesh out our initial utterance of the characterising phrase—no mental activity we can perform that will constitute the successful, as opposed to the unsuccessful, conceiving of a physiological zombie. However, as I noted at the start of this section, the force of this consideration is, to some degree, open to question. I will rest my case against the conceivability of zombies primarily on the problems raised in the previous section, therefore.

6. The Zombie Attack on Functionalism

So physiological zombies are inconceivable. They are inconceivable, primarily, since their mere possibility is inconsistent with something we know to be true—the falsity of phenomenal epiphenomenalism in the actual world. From this it follows that, if the actual world is causally closed under physics, then physicalism

is true with respect to consciousness: the facts about conscious experience supervene logically upon the physical facts in the actual world.

The arguments adduced above against the possibility of physiological zombies have had one factor in common: they all make use of the physiological zombist's commitment to phenomenal epiphenomenalism. But this commitment is required only for *physiological* zombies. The defender of *functional* zombies is faced with no such implication: the logical possibility of functional zombies requires that the presence or absence of phenomenal consciousness makes no *functional* difference, but it need not follow from this that consciousness makes no physical difference at a level lower than the functional level in question—that is, for any given functional role, at the 'level' of the occupant rather than the role. For example, the plumbing of a house satisfies a certain functional description—relating water flows, temperatures, pressures, and so on—and it might continue to do so even if, say, all its metal pipes were replaced with plastic ones of the same dimension. And of course the difference between metal and plastic is not epiphenomenal.[39]

Since, consistently with the logical possibility of functional zombies, consciousness need not be epiphenomenal in the actual world, then functional zombies are consistent with our actual ability to report our own consciousness and to denominate conscious states, and with the naturalistic emergence of consciousness during the history of life. For example, in the actual world our reports of consciousness might be caused appropriately by conscious states which

occupy certain functional roles, while in other possible worlds those self-same roles might be occupied by non-conscious states. In other words, the fact that a functional zombie may not issue genuine reports of its conscious experience does not entail that in the actual world, where we are not zombies, our reports are not genuine.

The arguments in the previous section directed at the positive conceivability of physiological zombies also fail to apply to functional zombies, for the simple reason that there *is* some third-person specifiable difference between my (merely) functional zombie twin and myself. Perhaps I can imagine my functional zombie twin by imagining a difference in our realizations (e.g. perhaps my twin is controlled by a radio link with a Universal Turing Machine located outside its body and running 'my' functional program at extremely high speed) … or at least the arguments presented above do not show I can't.

If functional zombies are to stand or fall, then, it will have to be for some other reason than the reasons for the inconceivability of physiological zombies. This is a point that is simple to make, but it is one of quite substantial significance. Zombie arguments are not monolithic, and refutations of one variety may be powerless against another; although we have established that physiological zombies are inconceivable, the possibility remains that functional zombies *are* appropriately conceivable, hence (assuming arguments from conceivability are cogent) logically possible, and hence that functionalism is false as a theory of phenomenal consciousness. Furthermore, the line of attack on

zombie-conceivability that has the most promise—the one that connects zombies with epiphenomenalism—will not work against functional zombies.

What, then, can we say about the conceivability of functional zombies—are functional zombies conceivable? The most plausible *prima facie* answer, overwhelmingly, is yes. Unlike the case with the physiological zombie, there is nothing conceptually incoherent in the notion of an organism identical in all its relevant[40] functional properties and states with me but entirely lacking the phenomenal life which I enjoy. Functional characterisations are, of course, multiply realizable, and notoriously can have many very non-standard realizations of the Chinese room / economy of Bolivia / meteor shower sort: it is, at a minimum, not a *conceptual* truth that highly non-standard realizations of precisely my functional architecture will have precisely my conscious inner life.[41]

Indeed, this very fact seems to be tacitly recognized by most contemporary defenders of functionalism: *analytic* functionalism, which would motivate an entailment from the functional to all aspects of the mental, is no longer a common position (and even in those days when it was popular—e.g. Armstrong (1968), Lewis (1972)—it was more usual to take the propositional attitudes as implicands rather than states of phenomenal consciousness). Instead, the prevailing assumption is that functionalism is empirically plausible and, if true, only contingently rather than analytically so. That is, it is not merely that consciousness might have been realized non-functionally, but that functionalism is only contingently true in the sense that the functional facts (in this world) do

not *conceptually entail* the phenomenal facts.[42]

But this resort to contingency leaves functionalism vulnerable to zombie attacks. If functionalism is only contingently true in this sense, then there is no conceptual incoherence in the supposition of functional zombies and so functional zombies are logically possible. This possibility—if zombie-style arguments are compelling in general—is all that is required to refute functionalism with respect to phenomenal consciousness.

My purpose here is not to establish flat-out that functionalism is false for consciousness but to emphasise the *asymmetry* between functionalism and materialism with respect to the zombist position. The materialist has the resources to show that physiological zombies are inconceivable, and hence to parry the zombist's claim that they are logically possible and halt the argument there. The functionalist does not have these resources and so must either deal with the zombist at a later stage of the argument[43] or succumb.

7. Conclusion: An Argument for Biological Theories of Consciousness

What I have argued in this paper is that zombie arguments suggest that phenomenal consciousness can be neither non-physical (assuming causal closure) nor computational/functional. This has two major morals. First, it means that zombie arguments fail to establish the falsity of physicalism. Second, if we take zombie arguments seriously and start from the assumption of causal closure, it follows that the best and only place to look for consciousness is in the

realm of the physical 'below' the level of the functional: consciousness must be a physical, rather than a functional or computational, attribute of collections of matter which relevantly resemble brains.

To put this latter point a little more precisely, since there is no one 'level' that is 'the' functional level but rather a hierarchy of occupant-role relations (see Lycan 1987), we should say that consciousness must be a more *biological* phenomenon than it is a computational one. That is, it must be sufficiently far down the occupant-role hierarchy in the brain that the possibility of different instantiations of those roles lacking consciousness ceases to be genuinely conceivable, preferably because we know that *whatever* occupies that role must be a realization of phenomenal consciousness in every possible world in which all the other relevant facts are held constant.[44]

References

Armstrong, D. 1968. *A Materialistic Theory of the Mind*, London: Routledge & Kegan Paul.

Balog, K. 1999. "Conceivability, Possibility, and the Mind-Body Problem." *Philosophical Review* 108 (1999): 497–528.

Block, N. 1980a. "Troubles with Functionalism." In *Readings in the Philosophy of Psychology, Volume 1*, ed. N. Block. Cambridge, MA: Harvard University Press: 268–305.

Block, N. 1980b. “Are Absent Qualia Impossible?” *Philosophical Review* 89 (1980): 257–274.

---. 1981. “Psychologism and Behaviorism.” *Philosophical Review* 90 (1981): 5–43.

Block, N., Flanagan, O., and Güzeldere, G. 1997. *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Block, N., and Stalnaker R. 1999. “Conceptual Analysis, Dualism, and the Explanatory Gap.” *Philosophical Review* 108 (1999): 1–46.

Braddon-Mitchell, D. 2003. “Qualia and Analytical Conditionals.” *The Journal of Philosophy* 100 (2003): 111–35.

Brueckner, A. 2001. “Chalmers’s Conceivability Argument for Dualism.” *Analysis* 61 (2001): 187–193.

Campbell, K. 1970. *Body and Mind*. London: Macmillan.

Carruthers, P. 2000. *Phenomenal Consciousness*. Cambridge: Cambridge University Press.

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

---. 1996. *The Conscious Mind*. New York: Oxford University Press.

---. 1999. “Materialism and the Metaphysics of Modality.” *Philosophy and Phenomenological Research* 59 (1999): 473–496.

---. 2002. “Does Conceivability Entail Possibility?” In *Conceivability and Possibility*, ed. T. Szabo Gendler and J. Hawthorne. New York: Oxford

University Press: 145–200.

---. 2003. “The Content and Epistemology of Phenomenal Belief.” In *Consciousness: New Philosophical Perspectives*, ed. Q. Smith and A. Jokic, Oxford: Oxford University Press: 220–272.

---. 2004. “Imagination, Indexicality, and Intensions.” *Philosophy and Phenomenological Research* 68 (2004): 182–190.

Cottrell, A. 1999. “Sniffing the Camembert: On the Conceivability of Zombies.” *Journal of Consciousness Studies* 6 (1999): 4–12.

Dietrich, E., and Gillies, A. 2001. “Consciousness and the Limits of Our Imagination.” *Synthèse* 126 (2001): 361–381.

Dretske, F. 2003. “How Do You Know You Are Not a Zombie?” In *Privileged Access: Philosophical Accounts of Self-Knowledge*, ed. B. Gertler. Burlington, VT: Ashgate Publishing: 1–14.

Elitzur, A. 1995. “Consciousness Can No More be Ignored.” *Journal of Consciousness Studies* 2 (1995): 353–358.

Flanagan, O. 1991. *The Science of the Mind*. Cambridge, MA: MIT Press.

Flanagan, O., and Polger, T. 1995. “Zombies and the Function of Consciousness.” *Journal of Consciousness Studies* 2 (1995): 312–372.

Gendler, T., and Hawthorne, J., eds. 2002. *Conceivability and Possibility*, New York: Oxford University Press.

Güzeldere, G. 1995. “Varieties of Zombiehood.” *Journal of Consciousness Studies* 2 (1995): 326–333.

Hawthorne, J. 2002. “Advice for Physicalists.” *Philosophical Studies* 109 (2002): 17–52.

Huxley, T.H. 1874. “On the Hypothesis That Animals Are Automata, And Its History.” *Fortnightly Review* 16 (1874): 555–580.

Jackson, F. 1986. “What Mary Didn’t Know.” *Journal of Philosophy* 83 (1986): 291–295.

James, W. 1879. “Are We Automata?” *Mind* 4 (1879): 1–22.

Kirk, R. 1974a. “Sentience and Behaviour.” *Mind* 83 (1974): 43–60.

---. 1974b. “Zombies v. Materialists.” *Aristotelian Society Supplementary Volume* 48 (1974): 135–152.

---. 1999. “Why There Couldn’t Be Zombies.” *Aristotelian Society Supplementary Volume* 73 (1999): 1–16.

---. 2003. “Zombies.” In *The Stanford Encyclopedia of Philosophy* (Fall 2003 Edition), ed. E.N. Zalta. URL = <http://plato.stanford.edu/archives/fall2003/entries/zombies/>.

Levine, J. 1998. “Conceivability and the Metaphysics of Mind.” *Noûs* 32 (1998): 449–480.

---. 2001. *Purple Haze: The Puzzle of Consciousness*. New York: Oxford University Press.

Lewis, D. 1972. “Psychophysical and Theoretical Identifications.” In *Readings in the Philosophy of Psychology, Volume 1*, ed. N. Block. Cambridge, MA: Harvard University Press: 207–215.

Lycan, W. 1987. *Consciousness*. Cambridge MA: MIT Press.

Marcus, E. 2004. "Why Zombies are Inconceivable." *Australasian Journal of Philosophy* 82 (2004): 477–490.

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

Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. Cambridge, MA: MIT Press.

Polger, T.W. 2000. "Zombies Explained." In *Dennett's Philosophy: A Comprehensive Assessment*, ed. D. Ross, A. Brook and D. Thompson. Cambridge, MA: MIT Press: 259–286.

Popper, K., and J. Eccles 1977. *The Self and its Brain*. New York: Springer-Verlag.

Shoemaker, S. 1975. "Functionalism and Qualia." *Philosophical Studies* 27 (1975): 291–315

---. 1981. "Absent Qualia are Impossible." *Philosophical Review* 90 (1981): 581–599.

---. 1999. "On David Chalmers's *The Conscious Mind*." *Philosophy and Phenomenological Research* 59 (1999): 439–444.

Stalnaker, R. 2002. "What is it Like to be a Zombie?" In *Conceivability and Possibility*, ed. T. Gendler and J. Hawthorne. New York: Oxford University Press: 385–400.

Stout, G. F. 1931. *Mind and Matter*. Cambridge: Cambridge University Press.

Yablo, S. 1993. “Is Conceivability a Guide to Possibility?” *Philosophy and Phenomenological Research* 53 (1993): 1–42.

NOTES

[1] Philosophical zombies were introduced in their modern form by Robert Kirk (1974a, 1974b; see also Campbell's 'imitation man' (1970)), though the kernel of the idea goes back at least to the debate about 'conscious automata' at the end of the nineteenth century (see, e.g., Huxley 1874, Stout 1931). The first flowering of zombie-related arguments occurred in the 1970s with the so-called 'absent qualia' attacks on functionalism (Block 1980a, 1980b, 1981; Shoemaker 1975, 1981 … the notion of a zombie is a special case of the absent qualia possibility). They have recently been once more at centre stage, spearheaded by David Chalmers' systematic construction of a zombie argument against physicalism (Chalmers 1996). The *Journal of Consciousness Studies* Volume 2, Issue 4 (1995)—a special issue on zombies—provides a representative sampling, but zombies have made steady appearances in the literature since then.

[2] See Güzeldere 1995 and Polger 2000 for useful zombie taxonomies.

[3] In fact, the zombie-types identified here purport to refute not only their particular target theory but all 'higher level' theories of mind as well; thus functional zombies also call into question behaviourism, and successful arguments from physiological zombies would refute functionalism as well as identity-theory.

[4] Many theorists (e.g. Carruthers 2000) are attracted to the view that a naturalistic theory of consciousness need only explain phenomenal consciousness in the *actual*, and nomologically closely related, worlds, and is not in the business

of giving necessary and sufficient conditions for the phenomenon. On this understanding, a naturalistic theory of consciousness—say, some version of functionalism—is quite compatible with merely possible occurrences of consciousness that come apart from the relevant kind of functioning in non-actual nomologically quite different or non-physical worlds. All of this, however, is irrelevant to zombie cases, properly construed. A minimum commitment for any explanation of consciousness, naturalistic or otherwise, is that it provide *sufficient* conditions for consciousness, within a given nomological context; that is, any adequate explanation for consciousness must involve the commitment, however restricted, to the obtaining of some relation of logical supervenience. Zombies are test cases for just this claim: they are stipulated in such a way that they satisfy all the conditions required by the test theory, *including* the relevant nomological context (e.g. the laws of physics of the actual world), yet they lack consciousness. If successful, therefore, zombie arguments show that the target theory does not provide sufficient conditions for phenomenal consciousness, and hence fails to explain its occurrence.

[5] Anthony Brueckner (2001) even goes so far as to call this the Standard Objection to zombie-like modal arguments.

[6] The literature on this has become quite extensive. Yablo 1993, Levine 1998, and Chalmers 2002 are good starting points. Hawthorne 2002 is an interesting high-level discussion of the issue.

[7] This connection was seen at the time of the introduction of zombies to

contemporary philosophy of mind (Kirk 1974b), but has been pressed most strongly by David Chalmers (1996).

[8] In fact the situation is a little more complex than this, as I explain below.

[9] See Chalmers 1996, 41–42 for a particularly eloquent expression of this position. The claim that the logical supervenience of all facts on the physical facts is sufficient for physicalism is sometimes disputed; for example Eric Dietrich and Anthony Gillies speculate about "a kind of dualism … where the physical and the nonphysical are logically tied together" (2001, 379–380). However, the line between this kind of 'dualism' and standard non-reductive physicalism is so thin as to be practically non-existent. Consider, for example, that on the species of 'dualism' that Dietrich and Gillies envisage the non-physical properties in question—though they certainly may be *different* properties than any physical property—are, in every possible world, brought into existence merely by the obtaining of the relevant physical facts: it is hard to see then how they can be metaphysically *over and above* these facts in the sense required to establish dualism (i.e. in any sense stronger than that in which economic facts, say, are distinct from physical facts).

[10] Thus, for example, that one can imagine a creature made of 'the same stuff' as me but lacking consciousness in worlds where the laws 'of physics' are different than in the actual world, is, even if true, simply irrelevant to the question of the truth of physicalism (and so, although this would perhaps be a species of zombie, it is not the type we are presently dealing with). The minimal thesis of

physicalism, as it is generally construed, is merely the doctrine that everything in the actual world is, or is logically supervenient on, the (type of) physical facts which are true of the actual world.

[11] Chalmers canonically defines a zombie world in this way; for example, he introduces the notion as "a world physically identical to ours, but in which there are no conscious experiences at all" (1996, 94).

[12] In a similar spirit, one might adopt formalist or Humean accounts of causation which make the mere law-like co-instantiation of conscious and physical events sufficient for causal laws. I will not discuss this possibility here; Chalmers discusses this kind of move, but cannot himself muster much enthusiasm for it (1996, 151 ff.).

[13] One might wonder if this would make *physics* quepiphenomenal also: this would be so only for some physical events, and only if phenomenal consciousness was sometimes by itself a *sufficient* cause for physical behaviours (a much stronger, and even less plausible, thesis than the mere thesis of overdetermination). It is also worth noting that conscious states that are quepiphenomenal in the actual world might be full-bloodedly epiphenomenal in other possible worlds, including some zombie-like worlds (where the physical laws are the same as those in the actual world but the laws of nature governing conscious causation are difference or absent): but this *mere possibility* of epiphenomenalism can—and does—do no work in my arguments in this paper.

[14] It is also worth noting, in passing, that consciousness might not be strictly

epiphenomenal if it had causal effects *on itself*, even though it had no effects on the physical (as is the case in certain species of parallelism); as with quepiphenomenalism, this caveat has no implications for the issues under discussion here.

[15] This is not, of course, to say that physicalism is *impossible*, in the sense that there could be no possible world of which physicalism is true. (Indeed, zombie worlds are supposed by the dualist to be just this sort of world.) Rather, it is to show that the truth of physicalism is logically inconsistent with what we know of *this* world, including what we know of both real-world physics and consciousness.

[16] I do not intend to be presupposing determinism here. In the case of indeterministic causation, such as quantum phenomena, what I mean by causal closure is that the probabilistic physical laws governing event-transitions would remain unchanged as long as everything physical were held constant. By contrast, an indeterministic system would not be causally closed under physics iff the event-transition probabilities could vary even when every element of the physical antecedent was fixed: if, for example, the collapse of the wave-function were brought about in some way by the intervention of (non-physical) consciousness, then the removal or varying of consciousness would radically change the quantum probabilities.

[17] See Chalmers 1996, 153–55. There is a way to understand panprotopsychism that makes it consistent with conscious epiphenomenalism: where the 'physical'

is the *relational* structure of the universe, *and* this can be held constant no matter how the intrinsic nature of things is varied. This may be Chalmers' own conception of the doctrine.

[18] John Perry (2001, Chapter 4) makes a similar point, though his concern there is to defend what he calls 'antecedent physicalism' and so his focus is importantly different: his central claim is that *given* that conscious states are physical states with causal powers *then* zombie-worlds are impossible.

[19] The claim I am making here—that the possibility of zombies entails the epiphenomenalism of consciousness in the actual world—is not always recognised. For example Robert Kirk has written that: "It is sometimes assumed that the view that zombies are possible entails epiphenomenalism; but that is not so. One may hold that zombies are possible while denying that the actual world is physically closed under causation: one might be an interactionist" (2003). However Kirk does not go on to defend this view, and I cannot see how it can be defended, unless this claim involves a tacit appeal to causal overdetermination; otherwise, the removal of consciousness would leave causal gaps and hence lead to changes in the physical history of the organism. As Kirk himself concedes, just a few lines later, "the zombie idea implies a conception of phenomenal consciousness on which it would be conceivable that a person's qualia should be stripped off like a jacket, leaving a fully functioning body. Given common assumptions, that would rule out causation by qualia in such a world" … and, according to the zombist, the possibility of physiological zombies entails that

phenomenal consciousness has this jacket-like character in the actual world.

[20] The actual Bush *could not* have a zombie twin, recall, because *ex hypothesi* removing consciousness from the actual Bush would change his physical behaviour and so there is no possible world containing an entity physically indiscernible from the actual Bush—the same physical configuration embedded in the same matrix of physical laws—but lacking consciousness.

[21] Take any world w which is physically as similar as possible to this one, in which the distribution of phenomenal consciousness is indiscernible from that in the actual world, and which is causally closed under physics. The actual world may or may not be able to stand in for w (depending on whether it is in fact causally closed or not). The zombist argues as follows. For any world w there is a twin zombie-world w_z that is indiscernible from w except that phenomenal consciousness is absent. This means that physicalism is false of w, since facts about consciousness are not fixed by the physical facts in w, and also that consciousness is epiphenomenal in w, since it is possible to remove consciousness—as in w_z—and leave everything physical unchanged. Our world may not be a member of the set $\boldsymbol{W}$ of worlds that can stand in for w; if it is not, physicalism is false and consciousness may not be epiphenomenal. But by the same token, if our world is not in the set $\boldsymbol{W}$ then it has no zombie twin, and so zombie arguments cannot work against it.

[22] These spirits could even be phenomenally conscious, consistently with the impossibility of zombies: the inconceivability of physiological zombies establishes

that physics is sufficient for phenomenal consciousness, not that it is necessary for it.

[23] For example, Chalmers writes that epiphenomenalism "is *only* counterintuitive, and … ultimately a degree of epiphenomenalism can be accepted" (1996, 151).

[24] In particular, some of these arguments have often become tangled up in debates about *self-knowledge* in ways that I think have obscured their import for physicalism, and I try to avoid that problem here.

[25] Avshalom Elitzur (1995) discusses this problem, which he calls the Bafflement Problem.

[26] Perhaps this should reduce the temptation to think that Mary's surprise, in Jackson's famous thought experiment (Jackson 1986), is any kind of *evidence* for the presence of phenomenal facts over and above the physical.

[27] Shoemaker 1999 makes a point similar to this, though he is discussing it in the context of the Paradox of Phenomenal Judgement (see below).

[28] Interestingly, to interpret zombies as referring to their own conscious states would be to convict them of massive and systematic error—since they have no such states—and so such an interpretation should be ruled out by the Principle of Charity; however, *from the point of view of an interpreter*, zombies are in principle indiscernible from normal people, and so must fall under exactly the same web of interpretation; thus, since zombies—if they are possible at all—cannot be interpreted as referring to internal phenomenal states, neither can we!

(This will be so unless the Principle of Charity can be supplemented with the non-empirical assumption that *this* is not a zombie world, but that seems hardly to be in the spirit of interpretationist semantics.)

[29] Katalin Balog (1999) makes her objection to conceivability arguments turn on the claim that when a zombie says "I am conscious" they are saying something that is not only meaningful but *true*. Chalmers' response has been to deny this (as in, e.g., his 2003). I need not adjudicate this dispute here, but note that Balog's argument suggests that if Chalmers were to abandon his allegiance to acquaintance-based semantics he would face problems on other fronts as well as the Paradox of Phenomenal Judgement.

[30] The appeal to evolution to show that epiphenomenalism is false is most often associated with Popper and Eccles 1977, but goes back at least to William James (1879). Chalmers recognises this consequence of epiphenomenalism: "The process of natural selection cannot distinguish between me and my zombie twin. … It follows that evolution alone cannot explain why conscious creatures rather than zombies evolved" (1996, 120).

[31] This is Chalmers' tactic (1996, 171). In response to the challenge of explaining the laws themselves, he writes that "beyond a certain point, there is no asking 'how'" (1996, 170).

[32] This section leaves open for the zombist, as the previous section did not, recourse to the weaker claim that a statement S is conceivable merely if ~S is not a conceptual truth. (See, for example, Balog 1999, 498–99.)

[33] See Cottrell 1999 for a thoughtful exploration of this problem for the zombist, and Marcus 2004 for an extended version of this argument and its consequences.

[34] It may also fall afoul of the need to provide criteria for transworld identity when considering counterfactual situations "about ourselves in which we do not have phenomenal experience," as Dietrich and Gillies 2001 argue—I will not deal with this point here.

[35] Chalmers distinguishes between introspective *beliefs*, which may perhaps be partially constituted by phenomenal states, and *judgements*, which are defined as "what is left of a belief after any associated phenomenal quality is subtracted" (1996, 174); *if* physiological zombies are possible, then judgements must be physically and so functionally identical with their corresponding beliefs. That is, removing the phenomenal must make no difference to the *cognitive relations* between a zombie's mental states (though, obviously, it may change the *nature* of these mental states, as they are experienced 'from the inside').

[36] This issue is usefully discussed by Joe Levine (2001, 159–167). Fred Dretske also argues that it is mysterious how we can know we are not zombies: he argues that "[t]here is nothing you are aware of, external or internal, that tells you that, unlike a zombie, you are aware of it. Or, indeed, aware of anything at all" (2003, 1).

[37] Note that this point does not depend on me not being able to know that I am not a zombie—perhaps if I believe that I am conscious for sufficiently good (though not conclusive) reasons, and it happens to be true, then I count as

knowing that I am not a zombie. The crucial point for this argument is that I cannot *know I know* that I am not a zombie—that I can't, from the inside, tell the difference between knowing I'm not a zombie and merely falsely believing I am not.

[38] Note that Kirk has since changed his views—see Kirk 1999.

[39] This example is taken from Polger 2000.

[40] This will include, naturally, not only my overall behavioural functions but also whatever internal functional structure our best functionalist theory requires.

[41] The voluminous inverted-spectrum literature on functionalism is also relevant here, insofar as it deals with phenomenal change while holding the functional constant: see Block et al. 1997, Section IX, for a starting point. The discussion of fading and dancing qualia, initiated by Chalmers 1996, suggests a tight empirical connection between cognition (understood functionally) and the structure of the phenomenal, but Chalmers argues that this correlation falls short of the logically necessary one that would be required—according to the zombist—in order to save functionalism.

[42] This is not to say that analytic functionalism is without recent defenders. David Braddon-Mitchell's "Qualia and Analytical Conditionals" (2003) directly addresses the issue of zombies for analytic functionalism; see also Hawthorne 2002 and Stalnaker 2002.

[43] This would require arguing either that functionalism does not involve commitment to the logical supervenience of the mental on the functional (i.e.,

that functionalism is true though fixing the functional is not logically sufficient for fixing the mental), or that conceivability is in general inadequate evidence for logical possibility (see, for example, Block and Stalnaker 1999 and the papers in Gendler and Hawthorne 2002).

[44] Thanks to two anonymous referees for very helpful comments on an earlier draft of this paper. This work was supported by a grant from the Research Enhancement Fund of the College of Arts, University of Guelph.