Materialism and the Metaphysics of Modality

DAVID J. CHALMERS
University of Arizona

Contents

1 Introduction
2 A priori entailment (Shoemaker)
3 A posteriori entailment
   3.1 The form of the argument (Hill & McLaughlin)
   3.2 Are there logically possible zombie worlds? (Yablo, Loar)
   3.3 Are there strong necessities? I. Examples (Yablo)
   3.4 Are there strong necessities? II. Explanations (Hill & McLaughlin, Loar)
   3.5 Modal rationalism
   3.6 Type-B materialism: Loose ends (Hill & McLaughlin, Loar)
4 Epiphenomenalism and the paradox of phenomenal judgment (Hill & McLaughlin, Shoemaker)
5 Dancing qualia (Shoemaker)

1 Introduction

The argument against materialism in The Conscious Mind has two parts. The first part, in Chapter 3 of the book, argues that there is no a priori entailment from physical truths to truths about consciousness. The second part, in Chapter 4 of the book, argues that there is no a posteriori necessary entailment from physical truths to truths about consciousness. Either part of the argument might be resisted. Corresponding to these paths of resistance, there are two very different brands of materialism, which I call type-A and type-B materialism.

Type-A materialists hold that phenomenal truths (insofar as there are such truths) are necessitated a priori by physical truths. Such a materialist denies that physically identical zombie worlds or inverted-qualia worlds are coherently conceivable, denies that Mary (of the black-and-white room) gains any factual knowledge on seeing red for the first time, and typically embraces a functional (or eliminative) analysis of consciousness.
Type-B materialists accept that phenomenal truths are not necessitated \textit{a priori} by physical truths, but hold that they are necessitated \textit{a posteriori} by physical truths. Such a materialist accepts that zombie worlds or inverted-qualia worlds (often both) are coherently conceivable but denies that such worlds are metaphysically possible, holds that the factual knowledge that Mary gains is knowledge of an old fact in a new way, and typically embraces an \textit{a posteriori} identification of consciousness with a physical or functional property.

All of the commentators in this symposium are type-B materialists, and three of the commentaries focus on my argument against an \textit{a posteriori} necessary entailment. This makes for a focused and productive discussion. I would like to thank all the commentators for their detailed, sophisticated, and generous commentaries, which in all cases significantly advance the discussion of important issues.

I will organize my replies to parallel the discussion in the book. In order, I will discuss issues relevant to the argument against \textit{a priori} entailment (Chapter 3), then those relevant to the argument against \textit{a posteriori} entailment (Chapter 4, backed by Chapter 2), then those relevant to epiphenomenalism and the paradox of phenomenal judgment (Chapter 5), and finally those relevant to the fading and dancing qualia arguments for nonreductive functionalism (Chapter 7). The first, third, and fourth of these parts mostly revolve around Shoemaker's commentary. The second part, and by far the longest, mostly revolves around the commentaries by Hill & McLaughlin, Loar, and Yablo.

\section{2 \textit{A priori} entailment}

All of the commentators are type-B materialists, holding that the phenomenal facts are not entailed \textit{a priori} by the physical facts, or do not logically supervene on those facts, as I put in the book. Unlike the others, Shoemaker endorses one element of the type-A materialist position, in that he denies that zombies are conceivable. He allows that inverted spectra are conceivable, which is enough to defeat logical supervenience in general. In effect, he holds that some phenomenal facts (facts concerning the existence of consciousness in a system, and facts about qualitative similarity) supervene logically on physical facts, but others (facts about the specific nature of experiences) do not. Facts in the second class are entailed only \textit{a posteriori} by physical facts.

Shoemaker does not argue directly for his position here. But he says that its coherence threatens my arguments in two places. First, he says it shows how there could be a necessary entailment from physical to phenomenal even without an \textit{a priori} entailment. Second, he says it shows that I am not "entitled" to the claim that zombies are conceivable.

Taking the second point first: Shoemaker is correct to say that the conceivability of zombies is not implied by the failure of \textit{a priori} entailment.
But he is incorrect in saying that I suppose that it is so implied, and in suggesting that my entitlement to the former rests on my entitlement to the latter. The direction of support is the other way around: the conceivability of zombies is used as one argument for the failure of logical supervenience. (There are other arguments for the failure of logical supervenience, and I note (p. 101) that one might endorse this failure without endorsing the conceivability of zombies.) The argument for the conceivability of zombies rests on quite distinct considerations.

Shoemaker may be misled by the sentence he quotes from Chapter 1 (p. 23): "If there were a functional analysis of the notion of experience or of phenomenal quality, then the analysis in question would yield functional analyses of specific phenomenal properties." I take the blame for misleading him here, as the sentence is ambiguous. If "the analysis in question" is taken to refer to the analysis in the first part of the sentence, this claim might be questionable. But the expression was in fact intended to refer to a quite different analysis under discussion earlier in that paragraph. The "analysis in question" is the analysis of specific phenomenal properties as "the sort of conscious experience that tends to accompany [functional property] P". If there were a functional analysis of the concept of conscious experience, one could plug it into the quoted analysis to yield a functional analysis of the specific phenomenal properties in question, as I asserted.

Shoemaker also suggests that his position shows how a psychophysical entailment can be necessary without being a priori. I don't think this is quite right, however. Even the conceivability of inverted worlds is enough for my arguments against a necessary connection (in Chapter 4) to go through. If we combine the conceivability of such worlds with the two-dimensional analysis of a posteriori necessity, it turns out that there can be no a posteriori necessary connection from physical facts to specific phenomenal facts (essentially because specific phenomenal concepts will have primary intensions that do not supervene on the physical), and Shoemaker's position will be ruled out. Of course Shoemaker may hold that there are a posteriori necessities that do not fit the two-dimensional model; but that point needs independent argument, and it is that argument that will do all the work. Merely exhibiting the position does no work, as my argument in Chapter 4 is itself an argument against the position.

Of course I disagree with Shoemaker's position on independent grounds. In particular I don't think that qualitative similarity and difference are functionally definable (partly because of the conceivability of zombies, and partly because I think that qualitative similarity depends only on intrinsic properties of experiences). Shoemaker says his claim should be "no more controversial" than my claim that judgment, memory, and other psychological states are functionally definable; but the claim about "judgment" and "psychological states" involve stipulative uses of those terms and are therefore trivial, while
the claim about memory is not made in any definitive way and plays no role in my arguments. Even if one accepts his claim, I don’t think it implies that specific qualia are fixed by physical facts, as he suggests. Shoemaker’s discussion of “realizing” functional definitions seems to suggest that specific qualia should be rigidly defined as the occupiers of the relevant functional role; but that definition would be incompatible with the failure of \textit{a priori} entailment. And without such a definition, there is no reason to think that fixing qualitative similarity fixes qualia. In any case, the failure of \textit{a priori} entailment (which Shoemaker concedes) is all that my central arguments require.

This touches on a minor misdescription of my argument in Chapter 3 that occurs repeatedly. Loar says my argument against \textit{a priori} entailment is “grounded” in the conceivability of zombies, and Yablo says “everything turns on” the claim that there are zombie worlds. In fact the zombie argument is just one of five arguments against \textit{a priori} entailment, and is not obviously the most important. First, the inverted spectrum and Mary arguments make claims that do not imply the zombie claim, so someone (like Shoemaker) who rejects zombies may still accept the first two. Second, it may be more correct to say the zombie argument is grounded in the argument from absence of structural/functional analysis than vice versa. Finally, at least dialectically, the Mary argument often works better than the zombie argument against a type-A materialist (a conceivability claim is easy to deny, a new-knowledge claim much harder). So I would not like to give the impression that everything depends on zombies here.

3 \textit{A posteriori} entailment

Type-B materialism has been perhaps the most popular position on consciousness in recent years. It simultaneously promises to take consciousness seriously (avoiding the deflationary excesses of type-A materialists) and to save materialism (avoiding the ontological excesses of the property dualist). For these purposes, Kripke’s introduction of \textit{a posteriori} necessities has seemed just what is needed. In recent years, almost every type-B materialist (Loar is a notable exception) has appealed to Kripkean examples in support of their position. On such a view, the relationship between consciousness and physical processes is supposed to be like that between water and H2O, or between Hesperus and Phosphorus.

In \textit{The Conscious Mind} I argue that this does not work. These (and other) standard examples of \textit{a posteriori} necessity give no support to a type-B materialist. Kripke’s \textit{a posteriori} necessities can all be accounted for within a two-dimensional framework, which has just one space of worlds (the conceivable worlds) and two intensions (primary and secondary) for terms and statements over those worlds. On this framework, the distinction between conceivability and possibility, and between epistemic and metaphysical necessity, falls at
the level of statements, not worlds. In particular, a posteriori necessities are statements with a contingent primary intension and a necessary secondary intension. It is relatively easy to see that necessities of this sort cannot save the materialist. So if a posteriori necessity is to save the materialist, it must be a different sort of a posteriori necessity, a sort not wholly explicable by the two-dimensional framework, and not supportable by appeal to the Kripkean examples.

Little of this is disputed by the current commentators. Instead most of them bite the bullet and argue that psychophysical necessities are different in kind from the Kripkean examples, and not explicable by the two-dimensional framework. Let us call these a posteriori necessities not explicable by the 2-D framework strong necessities. A strong necessity is an a posteriori necessity that does not have a contingent primary intension. I argue in the book that there is no reason to believe in strong necessities, and that they are highly problematic. In response, Hill & McLaughlin and Loar offer what they take to be accounts of the roots of these strong necessities, and Yablo gives reasons why he thinks strong necessities must exist. Before discussing these crucial counterarguments, I will first clarify some issues concerning my argument.

3.1 The form of the argument

Hill & McLaughlin and Loar focus on the four-step argument at the beginning of Chapter 4. This formulation does not really engage the type-B materialist, however, as it was intended to be accessible to readers ignorant of the purported distinction between conceptual and metaphysical necessity. The material most relevant to a type-B materialist can instead be found in section 2 of chapter 4, on objections from a posteriori necessity, and also in section 4 of chapter 2, which outlines the 2-D framework.

As far as the four-step argument is concerned, Loar is correct that the jump from (2) to (3) is the part that a type-B materialist should dispute. Hill & McLaughlin dispute (2) instead, but this rests on a misunderstanding of my use of “logical possibility”. On my usage, “logically possible” is near enough to interchangeable with “conceptually possible” (not “metaphysically possible”), and is tied by stipulation to conceivability. Without noting this usage, many passages and arguments will seem nonsensical.

Here is a crucial passage from p. 68 of the book:

Most importantly, none of the cases we have seen give reason to believe that any conceivable worlds are impossible. Any worries about the gap between conceivability and possibility apply at the level of statements, not worlds: either we use a statement to misdescribe a conceived world (as in the Kripkean case, and the second Goldbach case mentioned), or we claim that a

---

1. Throughout this discussion, “conceivability” should be understood as ideal conceivability, or conceivability on ideal rational reflection.
statement is conceivable without conceiving of a world at all (as in the first Goldbach case). So there seems to be no reason to deny that conceivability of a world implies possibility. I will henceforth take this for granted as a claim about logical possibility; any variety of possibility for which conceivability does not imply possibility will then be a narrower class. Someone might hold that there is a narrower variety of "metaphysically possible worlds"; but any reason for believing in such a class would have to be quite independent of the standard reasons I have considered here.

Perhaps it was not clear enough that "taking for granted" here is stipulative. I note, though, that the notion of logical possibility is introduced on pp. 34–35 with a tie to conceivability, and the quoted passage (and much of the rest of the book) does not really make sense otherwise. I also don't think my usage is unusual; it's typical in philosophy for the adjective "logical" to be tied to matters of conceptual truth and rational inference. I don't use "conceptual possibility" as the term doesn't seem quite right to me: it downplays the role of modal intuition (as opposed to mere semantic intuition); and while mathematical truths are certainly logically necessary on my usage, it is widely held that many of them are not conceptually necessary. But if a reader wants to substitute "conceptual" for "logical" on every occurrence, she will not be led too badly astray.

Of course, I hold that conceptual possibility = logical possibility = metaphysical possibility (at the level of worlds). But when we are discussing a potential distinction between conceptual and metaphysical possibility, "logical possibility" always goes with the former and not with the latter. Hill & McLaughlin read it as going with the latter, hence the confusion.²

(Two further clarifications: Hill & McLaughlin locate my "argument" for premise (2) in an incidental passage on p. 131, but in fact it is given in detail in Chapter 3. And they say that I hardly discuss the view that phenomenal states are a posteriori identical with neurobiological states, but in fact it is a special case of the type-B materialist view that I discuss at considerable length.)

With this matter clarified, the step that a type-B materialist must question is certainly the step from the failure of logical supervenience to the failure of metaphysical supervenience (and hence the falsity of materialism). My argument for this step has three parts. The first part (section 4 of chapter 2) outlines the two-dimensional account of a posteriori necessity and argues that it handles all standard a posteriori necessities without a gap between conceivable and possible worlds. The second part (section 2 of chapter 4) applies the 2-D account to argue that a posteriori necessities of the usual variety cannot save

² Hill's commentary on my book in Analysis (Hill 1998) rests largely on this misunderstanding. Hill says that my claim that we have a priori access to the space of logically possible worlds "begs the question"; but when "logical possibility" is understood appropriately, the claim becomes relatively uncontroversial. A brief online response to this and other articles can be found at http://ling.ucsc.edu/~chalmers/responses.html.
materialism. The third part (later in the same section) argues that there is no reason to believe in "strong necessities" which escape the 2-D account.³

The first two parts of the argument are straightforward, and none of the commentators dispute them. But a point of clarification. Hill & McLaughlin and Loar say that (in the second part) I rely on the Kripke-style observation that when it comes to phenomenal concepts, primary intension and secondary intension are the same. In fact (as Yablo notes), while I accept this observation, it is inessential to the argument. Even if primary and secondary inten­sions of phenomenal concepts differ, the argument against materialism goes through, although things have to be put slightly differently.

It is easy to see how, if we accept the 2-D account of a posteriori neces­sity, any apparent "holes" in my argument are filled in.⁴ Loar’s question about the step from distinct primary intensions to distinct properties is answered straightforwardly, for example: such intensions will provide different functions from worlds to extensions (remember, there is just one space of worlds), and therefore distinct properties (wrinkles about centering aside). Hill & McLaughlin question the same move, but get into unnecessary difficulties. The passage that they quote and criticize from p. 373 in fact ends as follows:

if the subject cannot know that R is P a priori, then reference to R and P is fixed in different ways and the reference-fixing intensions can come apart in certain conceivable situations. Unless we invoke the additional machinery of strong metaphysical necessity, the difference in primary intensions will correspond to a difference in reference-fixing properties.

Nothing Hill & McLaughlin say provides any reason to deny this. They change "reference-fixing intensions" in their quotation to "[properties]", but the change is not innocent. The claim in the first sentence involves primary intensions across conceivable worlds, and is relatively uncontroversial; even Loar accepts it. And the claim in the second sentence is that this difference in intensions yields a difference in properties unless one accepts strong necessi­ties (i.e., unless one holds that there are a posteriori necessities that escape the 2-D framework). Again, there is little reason to deny this.⁵

³ Yablo calls the view that all a posteriori necessities fall under the two-dimensional account “textbook Kripkeanism.” I think this may be the wrong term, partly because it is not clear that Kripke accepts the view himself, and partly because many textbook uses of a posteriori necessity are incompatible with it. If one needs a name for the position, I might suggest “two-dimensionalism” or perhaps “modal rationalism.”

⁴ A crisp formalization of the argument is in Chalmers 1998 (lecture 1). There is one loophole that remains. One can appeal to a distinction between primary and secondary intensions for fundamental physical concepts, postulating phenomenal or protophenomenal grounds for basic physical dispositions (TCM pp. 134–36). It is not entirely clear that the resulting “panprotopsychism” counts as materialism, but in any case it will be much more to my tastes than to those of most materialists.

⁵ Note that my claims in this passage require none of Hill & McLaughlin’s premises (a), (b), and (c). Hill & McLaughlin think I need (a) because they overlook the significance of their substitution; they think I need (b) and (c) because they overlook the significance
The residual question, then, is whether there are strong psychophysical necessities. Standard *a posteriori* necessities have a necessary secondary intension but a contingent primary intension. A strong necessity, by contrast, is an *a posteriori* necessity with a metaphysically necessary primary intension. The primary intension of a strong necessity (unlike that of a standard *a posteriori* necessities) will be false in no metaphysically possible world. Nevertheless, it seems that the primary intension will be false in some *conceivable* world, because the statement is *a posteriori*. Hence the crucial feature of strong necessities, which is that they seem to require a distinction between conceivability and possibility at the level of worlds.

In the book (p. 136), I call these “strong metaphysical necessities”, and characterize them by the view that there are fewer metaphysically possible worlds than logically possible worlds. Not all of the commentators recognize their view in this description. Hill and McLaughlin’s resistance is explained by their misreading of my use of “logically possible”. (They also reject my characterization of the view as requiring that such necessities stem from factors independent of the semantics of the terms involved. I should have said “independent of the two-dimensionality of the semantics of the terms involved”.) Yablo resists by denying that zombie scenarios involve “worlds” at all. Before addressing the question of whether there are strong necessities, I will address this point of Yablo’s.

### 3.2 Are there logically possible zombie worlds?

In an interesting line of resistance, Yablo accepts that it is logically (or conceptually) possible that there are zombies, but denies that there are logically possible zombie worlds. One way to make this denial would be to stipulate that “world” means “metaphysically possible world”. This would make the denial a mere terminological variant of the usual core issue: instead of discussing a gap between conceptually possible and metaphysically possible worlds, we would have to discuss a gap between conceptually possible “scenarios” and metaphysically possible worlds, where scenarios are world-like except that they are metaphysically impossible. But I think Yablo is making a deeper point, questioning whether there are any world-like objects that correspond to “zombie worlds” at all.

I think the answer to this question is clearly yes. If zombies are ideally conceivable, we need logically possible zombie worlds.\(^6\)

---

\(^6\) The same goes for inverted worlds. Here and in what follows, I use zombie worlds to stand in for any world physically the same as and phenomenally different from our own. If we allow that the primary and secondary intensions of phenomenal concepts may differ, this should be taken to include worlds in which the primary intensions of our phenomenal concepts yield different results.

---

of Loar’s term “introduce” in the quoted passage. As for my overall argument, I do require (a), but (b) and (c) are inessential.
To consider the issue, pretend for a moment that zombies are conceivable but not metaphysically possible. Even better, take another “strong necessity” view, on which the only possible worlds are those whose laws of nature are our laws. (This view differs from a weaker view on which apparently counternomic worlds should be redescribed as not containing “mass” and such at all. The latter view also endorses the metaphysical necessity of laws, but is compatible with the 2-D framework.) On such a view, counternomic scenarios, although conceivable, are reflected in no metaphysically possible worlds at all. But we still need logically possible worlds with different laws, for all sorts of reasons.

Think of the reasons why “possible worlds” talk is introduced into philosophy in the first place. Possible worlds are introduced to deal with counterfactual thought, the semantics of counterfactual language, rational inference, and the contents of belief, among other reasons. A scientist can think counterfactually (and rationally) about scenarios with different laws, and can make true utterances about these scenarios. If we are to use possible-worlds talk to characterize the contents of her beliefs in discovering laws, we will need to appeal to counternomic worlds. Without counternomic worlds, we will not be able to use worlds to make sense of her inference processes. And so on. Ruling out counternomic worlds will make possible worlds useless for many or most standard purposes. Even if someone insists that such worlds are not metaphysically possible, we need logically possible counternomic worlds.

Something similar applies to zombie worlds. Even on a type-B materialist view, we can think counterfactually (and rationally) about the possibility of a different distribution of phenomenal properties with the same physical properties. We need worlds corresponding to these possibilities to make sense of counterfactual thought, of the semantics of counterfactual utterances, of rational inferences involving consciousness, of the contents of rational beliefs about consciousness, and so on. We can write coherent science fiction about zombies, and speak coherently about the truth in such fictions. Talk of logically possible zombie worlds is justified in the usual way by their role in these uses.

I suppose one might resist by holding an extreme modal-realist view (even stronger than David Lewis’s) on which possible worlds are simply “out there” and on which their connection to the rational uses in question is neither here nor there. But even on such a view, we would end up having to postulate worldlike objects (“ersatz” worlds, at the very least) for the rational purposes. And on the dominant view on which all counternomic worlds are regarded as ersatz abstract objects, there seems no ground for resistance. It is easy to construct an ersatz object that behaves in just the way that a zombie world should. The obvious strategy is to use maximal consistent worldbooks, where “consistent” is understood in the a priori sense. One has to tread care-
fully in handling two-dimensional phenomena and centering, but the matter is straightforward.

Another line of resistance might liken zombie worlds to supposed "impossible worlds" with all their problems, but the analogy doesn't work. Impossible worlds simply don't behave as worlds should: statements are both true and false there, for example. In zombie worlds, statements can be semantically evaluated with ease, and they never come out both true and false. Zombie worlds can even be considered both as actual and as counterfactual, as a world should be, and statements will be well-behaved on both methods of evaluation.

Loar suggests a concern: if zombie worlds are logically possible but not metaphysically possible, such worlds won't satisfy the necessity of identity. I think this is a better argument against strong necessities than against zombie worlds; what causes problems is not the zombie worlds themselves but the postulation of strong necessities and the associated problematic identities. But if one does hold the strong necessity view, I think the right thing to say is that the necessity of identity applies only across metaphysically possible worlds, not logically possible worlds. At worst, if someone insists both on strong necessities and that worlds must satisfy the necessity of identity, we'll just have to call the zombie situation a "scenario" or some such instead. Nothing crucial changes.

So, it seems that there are overwhelming grounds for admitting something like zombie worlds into our ontology, at least given that zombies are ideally conceivable and that we have worlds in our ontology at all. More generally, there are overwhelming grounds for admitting a logically possible world for every ideally conceivable scenario: we need to do so in order to make sense of most of the things that possible worlds are supposed to make sense of. The only question is whether we also need to stipulate a distinct modality of metaphysical possibility, holding that only some of these worlds are metaphysically possible. I think not, but more on this shortly.

I have in effect just argued for Yablo's CP: we need a logically possible world for every ideally conceivable scenario if we are to make sense of the various rational notions in question. Yablo's (1)–(7) argument for CP is not my argument: on my reading, his (2) is weaker than his (1) and his (5), for example. Yablo is right, though, that I need to be able to move from de dicto to de re possibility where logically possible worlds are concerned, and that I need to exclude opacity phenomena. The conception of logically possible worlds I am using does just that. With a world for every ideally conceivable scenario, it follows that whenever it is conceivable that $P$, there will be a logically possible world in which $P$ (according to primary intension). And whenever two conceived-of scenarios are not identifiable even under ideal rational reflection, they will correspond to distinct logically possible worlds.
Yablo asks how we conceive of worlds in order to evaluate them under primary intensions. I take Yablo’s first option: we conceive of them under descriptions. I certainly don’t endorse the “immaculate conception” option. Perhaps a couple of my remarks suggest the “what I would say if plopped down in W” option; this may be a useful heuristic but it is not my considered view, partly for the reasons Yablo mentions. Instead, I take a (centered) world—always under a description, of course—and consider it as an epistemic possibility, asking “what is the correct thing to say if this world is actual”.

I avoid opacity phenomena not by conceiving under canonical descriptions (as Yablo suggests in his discussion of the first option), but simply because rationally inequivalent descriptions will pick out inequivalent worlds, by definition of logically possible worlds. Of course a world can be described in many different ways, but such descriptions will always yield the same results when primary intensions are evaluated, as the equivalence between the descriptions will be revealed on ideal rational reflection.

This applies at least when the descriptions under which we conceive of worlds are intended as “primary” descriptions. We can also conceive of worlds under “secondary” descriptions, if we like, but then all bets are off. (This is more or less the distinction between 1-conceivability and 2-conceivability outlined in the book, or Yablo’s distinction between conceiving of an E-verifying world and an E-satisfying world.) When worlds are conceived of in the secondary way, opacity phenomena can arise: we won’t in general know whether two descriptions describe the same world (unless we have relevant empirical knowledge of the actual world, or unless the concepts’ primary and secondary intensions coincide), and we can’t always know whether we have conceived of a world. When worlds are conceived of in the primary way, however, no such problems arise.

3.3 Are there strong necessities? I. Examples

The discussion of strong necessities in my book is relatively brief, as few people in the literature had embraced them; the great majority of type-B materialists appealed to the Kripkean examples instead. But I argue against strong necessities (pp. 136–38), saying there is no reason to believe in them. In particular, I say that (a) they cannot be supported by analogy with other a posteriori necessities; (b) they involve a far more radical sort of a posteriori necessity than Kripke’s, requiring a distinction between logical and metaphysical possibility at the level of worlds; (c) they lead to an ad hoc proliferation of modalities, (d) they raise deep questions of coherence, (e) strong necessities will be brute and inexplicable, and (f) the only motivation to postulate such necessities is the desire to save materialism. I stand by these points, but there is more to say about all of them, and each of the commentaries raises interesting questions.
The first part of the argument is the absence of analogy: there are no other counterexamples to the two-dimensional account of \textit{a posteriori} necessity. I think it is relatively uncontroversial that most \textit{a posteriori} necessities fall under the 2-D model, and certainly that the most commonly accepted examples do. Loar accepts explicitly that there are no other counterexamples to the 2-D model: he thinks that psychophysical necessities are \textit{sui generis}. Hill & McLaughlin appear to accept the same thing. Shoemaker touches on one source of counterexamples, though, and Yablo discusses a potential counter-example in detail.

Shoemaker mentions his view that the laws of nature are metaphysically necessary. As we have seen, this view can be interpreted as requiring that laws are strong necessities. These purported necessities are at least as tendentious as psychophysical necessities (and far less widely accepted), however, and I would use the 2-D framework to argue equally against them.

The same goes for the claim that God exists necessarily, which would also provide a strong necessity if true. Yablo adapts the God example to provide an ingenious argument for strong necessities, however. In effect, he notes that one of the following statements must be necessary: "a necessarily existing god exists", and "a necessarily existing god does not exist". But each of these statements is conceivably false (he suggests), and therefore \textit{a posteriori}; and the necessary \textit{a posteriori} status of either statement is not plausibly explicable by the 2-D framework. So one of these statements is a strong necessity.

I deny this, because I deny that a necessarily existing god is conceivable. Such a god may be conceivable in the sense of "not obviously inconceivable", but in no stronger sense. I certainly can form no clear and distinct conception of such a god (like many, I was suspicious of the idea the moment I heard about it as a student), and continued rational reflection reveals all sorts of problems with the idea. One thing that rational reflection reveals is the conceivability of a world \textit{without} a god. If such an intuition is accepted, it causes grave difficulty for the coherence of the idea of a necessarily existing god. (Ideal reflection allows us to use one modal intuition in evaluating others, of course.)

The problematic issues here arise because of the double modality: we are conceiving not just of worlds "in themselves", but also of what is possible or necessary within those worlds. Conceiving of a god (an omnipotent, omniscient, and benevolent being, say) is arguably not too hard; but to conceive in addition that the being exists necessarily, we have to conceive that the space of possible worlds is such that this god exists in each of them, \textit{despite} the conceivability of a godless world. That is, we have to conceive that the conceivability/possibility thesis (CPT: all conceivable worlds are possible worlds) is itself false. This is what does all the work in the example: if it is
conceivable that CPT is false, then (by CPT!) it is possible that CPT is false, and CPT is therefore false (as it is surely necessarily true, if true).

A natural response to this sort of argument is to restrict the conceivability/possibility thesis to claims about the distribution of nonmodal properties within worlds, leaving double modals outside its scope. I think this response would be defensible, and not entirely ad hoc (CPT might still apply to worlds in themselves, although not to "cosmoses" of possible worlds; all that changes is that one can’t make Yablo’s "de dicto to de re" step for double modals). But I prefer to hold onto the stronger thesis, by denying that it is conceivable that CPT is false. I hold that CPT is a priori, although highly nontrivial, like many theses in philosophy. In fact I will sketch an a priori argument for CPT later in this paper. If this is correct, then CPT is not conceivably false on ideal rational reflection, and it is not ideally conceivable that a necessarily existing god exists.

One might resist the conclusion by stripping down one’s conception of a god so that a god’s nonexistence becomes inconceivable. If we conceive of a god (nonrigidly) as “all there is”, for example, then a god exists necessarily. (My discussion above assumes that a god is required to have contingent-seeming properties such as omnipotence.) The case of numbers and sets, which Yablo raises, is something like this. I think it is a priori that numbers exist, so inconceivable that they do not exist. Of course this depends on one’s conception of numbers. There are alternative conceptions on which it is a posteriori that numbers exist, or on which it is a priori that numbers do not exist. What matters is that there is no tenable conception of numbers on which (a) there is a conceivable world in which numbers exist, (b) there is a conceivable world in which numbers do not exist, and (c) if numbers exist, they exist necessarily. I think there is no reason to deny this.7

3.4 Are there strong necessities? II. Explanations

If I am right so far, there are no good reasons to believe in strong necessities outside the psychophysical case. But Hill & McLaughlin and Loar both provide interesting potential explanations of why there should be strong necessi-

---

7 Another class of candidates for strong necessities might include mathematical statements that are necessarily true (even by primary intension) but not knowable a priori. It is far from clear that such statements exist (idealizing away from our cognitive limitations), but even if they exist, I do not count them as strong necessities. Unlike strong necessities, such statements open no gap between the conceivability and possibility of worlds, as opposed to statements; there is no conceivable world (considered as actual) in which these statements are false. I think such statements, if they exist, are best regarded as neither a priori nor a posteriori, but as a third category, "inscrutable". (The epistemic theory of vagueness, if true, would supply further inscrutable truths.) The existence of inscrutable truths would complicate some of the present issues, but they would not help the type-B materialist, who accepts that zombie scenarios are clearly conceivable. So I set them aside here; I discuss this topic further in Chalmers 1998 (lecture 2) and in forthcoming work.
ties in the psychophysical case. If this worked, it would be as good as a resolu-
tion of the mind-body problem as any. Unfortunately, I don’t think the proffered explanations work. Below I explain why.

Hill & McLaughlin offer a psychological explanation of why we can con-
ceive of zombies, in terms of the independent cognitive processes by which we conceive of physical processes and by which we conceive of experiences. There are some questions about the details of this explanation; and I should note that it is not obviously a purely physical/functional explanation, as it helps itself to the unreduced notion of a sensory concept. But let me grant that some explanation along these lines can be given. The question is whether it provides any explanation of strong necessities. On the face of it, it seems not. After all, one can also give a psychological explanation of why we can conceive of red squares, in terms of the distinct cognitive processes and epistemic constraints involved in conceiving of color and shape. One can give a psychological explanation of why we can conceive of five-horned ani-

dals, or of silicon-based life. But no-one would infer that there are strong necessities denying the metaphysical possibility of red squares or five-horned animals or silicon-based life.

The form of Hill & McLaughlin’s argument is far from unique to the psy-
chophysical domain. The precise details (concerning sensory concepts and recognitional concepts) may be unique, but these details are inessential. What does the work in their argument is the distinctness of the “faculties” or the “epistemic constraints” that these concepts involve. One could make an argument with the same form about shape concepts and color concepts, an argument that should then be just as compelling. And one could make a very similar argument concerning any two pairs of concepts with different cogni-
tive roles, thus “explaining away” any conceivability intuition at all. (There will always be a cognitive explanation of a modal intuition!) If this account tells us that Cartesian modal intuitions are unreliable, the same will go for all modal intuitions.

(Worse: a similar account might tell us that all of our mathematical beliefs are unreliable. There will presumably be cognitive explanations of these beliefs, too.)

The trouble is that an explanation of a strong necessity has to do two things: it has to show us why a state-of-affairs should be conceivable while at the same time being impossible. To put matters differently, it should explain why conceivability is an unreliable guide to possibility of such states-of-affairs (Hill & McLaughlin put it this way themselves). But Hill & McLaughlin have discharged only half of the burden, and the easy half: they have explained why zombies might be conceivable. But they have done noth-
ing to explain why we should take this conceivability to be unreliable.

Hill (1997) likens his strategy to Kripke’s “explaining away” of modal intuitions about heat and molecular motion, but I think it is crucially differ-
Kripke's strategy explains why necessity and apparent contingency are co-present: if we mistakenly describe a conceived situation by considering it as actual rather than as counterfactual, we should expect that certain necessary statements will falsely appear to be contingent. By contrast, Hill's strategy merely explains the appearance of contingency; why this appearance is co-present with necessity is not explained at all. At most, Hill tacks the hypothesis of necessity onto the end of his explanation of contingency, so it remains an unexplained explainer.

The closest Hill & McLaughlin come to addressing this burden is saying: "Given these differences between sensory concepts and physical concepts, a sensory state and its nomologically correlated brain state would seem contingently connected even if they were necessarily one" (my italics). But this may well be a deeply "per impossibile" counterfactual ("given mathematical concepts, 1+1 would seem to be 2 even if it were 3"). What we need is an explanation of how the two states could be necessarily one. Or ascending to the level of concepts, we need an explanation of how two such distinct concepts could pick out the same property, and indeed have the same property as a reference-fixer.

That is just what Loar attempts to provide. Some parts of Loar's discussion appeal only to the distinctness of cognitive role of the two sorts of concepts; as an "explanation", this has the same problems as Hill & McLaughlin's. But other parts attempt something more. In effect, Loar attempts to explain strong necessities by explaining how it is that physical and phenomenal concepts could pick out the same property and have the same property as their reference-fixer, despite their cognitive distinctness. The details of Loar's account here are somewhat vague, but they are spelled out in much more detail in Loar (1997), which I rely on below.

Loar appeals to two facts about phenomenal concepts: they are (a) recognitional concepts that (b) express the same property that they refer to. (In Loar's terminology, the property a concept "expresses" is its reference-fixer, or the property corresponding to its primary intension across possible worlds.) The significance of (a) is that recognitional concepts and physical-theoretical concepts are generally cognitively distinct even when they refer to the same property. But this is not enough to save materialism: other recognitional concepts conceive of their referents under contingent modes of presentation (i.e., they have distinct primary and secondary intensions across possible worlds), so they express a property quite distinct from that expressed by the corresponding theoretical property. So Loar adds (b), noting that in the phenomenal case, uniquely, we have recognitional concepts that express the property they refer to. He appeals here to Kripke's observation about pain, in effect noting that phenomenal concepts have the same primary and secondary intension, and do not conceive of their referents under contingent modes of presentation. Granting all this, we deduce that the two concepts are cogni-
tively distinct and coreferential (because of (a)), and that each concept expresses the property it refers to (because of (b)), from which it follows that the two concepts have the same reference-fixing property. So strong psychophysical necessities are explained.

The problem with this is straightforward. The introduction of point (b) undercuts the value of point (a). The significance of point (a) for Loar lies not just in the cognitive distinctness but also the coreference of recognitional and theoretical concepts. But this coreference is explained by the two-dimensional nature of such recognitional concepts: they typically conceive of their referent as "the cause of such-and-such experience", or under some similar contingent mode of presentation. If we remove this feature of recognitional concepts (as we do in accepting (b)), we no longer have any reason to believe that recognitional concepts and distinct theoretical concepts should corefer.

As things stand, in accepting (a) and (b) we are left with the observation that phenomenal concepts and physical concepts (i) are cognitively distinct, and (ii) both express the property that they refer to. It's clear that nothing here begins to justify the coreference of phenomenal and physical concepts. In fact the situation is the opposite: in every other case of concepts satisfying (i) and (ii), they have distinct referents. One might suppose that recognitional is doing some extra work here (thus distinguishing this case from other cases involving nonrecognitional concepts), but the only work it does for Loar is in explaining (i) and in providing cases of coreference when (ii) is false. So once (i) and (ii) are granted, there is nothing in Loar's account to justify coreference, and his explanation of strong necessity fails.

Loar slides over this point repeatedly. His argument often seems to have the form (1) Given that phenomenal concepts and physical-theoretical concepts corefer (as is typical with recognitional concepts), and (2) given that phenomenal concepts express the property that they refer to (as is unusual with recognitional concepts), then (3) phenomenal concepts will express physical-theoretical properties. This argument is valid, but Loar fails to note that premise (2) undercuts the support for premise (1).

A final way to put the problem is this. Loar's account requires, and rests on the claim, that a phenomenal concept ("feels like this") might conceive of a physical property under a necessary mode of presentation. This in turn rests on the claim that a physical property might have this mode of presentation

---

8 I am not sure that phenomenal concepts should be thought of as recognitional concepts, but I will go along with Loar on this point here. In his paper in this symposium, Loar sometimes says "experiential" rather than "recognitional", but he means something similar. "Experiential concept" has a different meaning for me (it is more or less synonymous with "phenomenal concept"), so I will use "recognitional" here.

9 This also requires the assumption (implicit in Loar's discussion) that a physical-theoretical concept expresses the same property that it refers to. I think this is arguable, but I will go along with it to keep things simple. If this assumption is retracted, related versions of Loar's point and my reply can still be given.
necessarily; that is, it rests on the claim that there might be a necessary connection between a physical property and a phenomenal feel. But this is just the sort of strong necessity that was in question and that we were trying to explain. So rather than explaining strong psychophysical necessities, Loar’s account presupposes strong psychophysical necessities.

We see, then, that neither Hill & McLaughlin’s nor Loar’s account can explain strong psychophysical necessities. Instead, both accounts assume such necessities at a key point, as a kind of primitive. Such necessities remain as mysterious and problematic as ever.

3.5 Modal rationalism

Hill & McLaughlin suggest that defeating all other candidates for strong necessities doesn’t count for much; all that it does is remove a few “counterexamples” to the two-dimensional analysis, or to the conceivability/possibility thesis. I think this analysis gets the dialectic wrong: most type-B materialists (even Hill 1991) support their case by appealing to these examples; and the reason that many are skeptical that conceivability implies possibility is that they think there are clear counterexamples to the thesis. So removing these counterexamples removes the central plank of support from the type-B position. And the failure of any apparent explanation of strong necessities weakens the position further.

Still, the position remains at least formally open. An opponent could hold that there are strong psychophysical necessities quite unlike the usual a posteriori necessities, that the space of metaphysically possible worlds is smaller than the space of logically possible (or conceivable) worlds, and that zombie worlds are excluded. If my arguments have been successful so far, I have removed many of the positive reasons for believing this. But perhaps the worry is that I have not shown why it couldn’t be true. So, why must it be false?

That is a deep and interesting question on which I hope to write at length in the future, but here I will say just a few words. Of course there are the further considerations against strong necessities that I raise in the book: they will be brute and inexplicable, that they lead to an ad hoc proliferation of modalities, and that they will be epistemically just like fundamental laws. And their introduction would lead to a philosophical revolution far more radical than Kripke’s. But an opponent might bite the bullet on all these points. What is fundamentally wrong with the idea?

The fundamental problem with the idea, I think, is that it rests on a false conception of modality. It is widely acknowledged that there is a circle of modal notions: possibility, consistency, and rational entailment, for example. Some of these, and perhaps the most grounded of all, are clearly rational notions, such as consistency and rational entailment. The framework of pos-
sible worlds is tremendously valuable in making sense of these notions; and as I suggested earlier, the space of worlds that we need here is the space of logically possible worlds, with one world for every ideally conceivable scenario. A narrower class of worlds is no help in making sense of these notions; breaking the link between conceivability and possibility breaks the link between rationality and modality.

Some might think that we cannot “break out of the circle” from rational notions to such “metaphysical” matters as \textit{a posteriori} necessity and the concept/property distinction, but that is wrong. The point of the two-dimensional analysis of \textit{a posteriori} necessity is that it allows us to explain such matters with only one modal primitive (the space of logically possible worlds). Given this primitive, we need only the two-dimensional intensional structure of our concepts (an \textit{a priori} matter of conceptual analysis) along with nonmodal facts about the actual world to ground all these matters of metaphysical modality.

So one modal primitive (plus conceptual analysis plus nonmodal facts) gives us everything. And it must be a primitive tied constitutively to such rational notions as consistency, entailment, and ideal conceivability. For if it is not tied constitutively to those notions, we will not be able to explain those notions (nonmodal facts certainly won’t help us here), and such notions are part of what a theory of modality needs to handle. If our choice of primitive is a space of worlds, it is clearly the logically possible worlds that we need; and if our choice of primitive is not a space of worlds, it will be some other notion in the rational/modal circle.

Advocate of strong necessities must reject this picture. They cannot reject a rational modality altogether, as they use such modal notions as consistency, rational entailment, and conceivability themselves (their position is partly defined with respect to these notions). So they must accept something akin to the space of logically possible worlds, although they might use another name. But they think there is a further metaphysical modality, and that not every logically possible world is a metaphysically possible world. This modality is not to be defined in terms of the rational modality, or even in terms of the rational modality plus nonmodal facts; it is a further primitive, part of a different circle entirely.

This picture is modal dualism. It requires two modal primitives: there is a space of logically possible worlds, and then (as a further primitive fact) a smaller space of these picked out as “metaphysically possible”. The primitives need not be cast in these terms, but we know that two primitives will be needed, as we need one to account for the rational modalities, and we need another in turn to account for the “metaphysical”.

Once we get this far, it is clear that something has gone wrong. There is no reason to believe in any more than one modal primitive, and there is no reason to postulate a second “metaphysical” primitive at all. The second
primitive is an invention; nothing in our conceptual system requires it. In particular, we need it to account for none of our existing modal notions. So it is a primitive that answers to no-one and does no work.

It seems to me that we do not even have a distinct concept of metaphysical necessity to which the second primitive can answer. The momentary impression of such a concept may be a residue from initial impressions of the Kripkean distinction between epistemic and metaphysical modality. But once we recognize that this distinction can be explained with one modal primitive, and that there are constitutive ties between the Kripkean modalities, the grounds for this impression disappear. The only concept of a "metaphysically possible world" that we have is that of a logically possible world. If someone thinks they have a distinct concept here, there is no reason to believe that anything answers to it.\(^\text{10}\)

I think there is overwhelming reason to reject modal dualism and accept only one modal primitive. If so, strong necessities are ruled out, and type-B materialism is false.

3.6 Type-B materialism: Loose ends

(1) Hill & McLaughlin say at one point that the counterintuitive nature of my conclusion is itself a flaw in my treatment of modality. I think this is making the counterintuitiveness do double duty. To be sure, the conclusions give reasons to resist the argument, and to look long and hard at it, but they are doing that already. The argument itself must be assessed on its merits. Hill & McLaughlin also suggest that my view is "wishful thinking", but I think the situation is precisely the opposite. If it were not that the antecedent impulse to believe materialism were so strong (I share it, too), and its denial so hard to accept, I think the arguments would be relatively uncontroversial. As things stand we are in the situation where we scratch around for a materialist "way out", and the existence of any potential loophole in the argument, no matter how ad hoc, is seized upon.\(^\text{11}\) I don't say this is a bad strategy, but I think it is clear where the wishful thinking lies.

(2) Loar mentions the explanatory gap. I think this is indeed a problem for the type-B materialist view (how could a physical process necessitate con-
conscious experience? why couldn’t God have created a zombie world?) and I think Loar’s answer doesn’t work, for the reasons I’ve given. But even if a type-B materialism is accepted, the explanatory gap is still a major issue. Loar treats the explanatory gap as if the only issue is whether it threatens materialism, and here he is like many philosophers who take the truth of materialism to be the major question in this whole area. But this is not my view. There are questions about explanation, and there are questions about ontology, and the former are as at least as important as the latter. In the book, I devote at least as much space to the explanatory issues, and in the end they may matter more in developing a theory of consciousness.

Although a type-B materialism will reject my conclusions about ontology, it will (in effect) accept my conclusions about explanation. It remains the case that crossing the gap requires epistemically primitive bridging principles. These principles will be called “identities” or “necessities” rather than “laws”, but their role in a theory will be much the same. A theory of consciousness will still need to have a primitive vertical structure, not derivable from the horizontal structure of physical theory. No matter how we interpret the ontology, the shape of a theory of consciousness will be entirely different from theories in other domains.

So even if one saves the letter of materialism, it is not clear that one saves the spirit. The materialist dream of a seamless explanatory web from physics on up will fail, and we will need to search for a theoretically independent bridge. We will eventually want to systematize and simplify this bridge until we are left with a set of simple (fundamental?) identities or necessities from which all the others follow. The result will be something much like the sort of psychophysical theory that I advocate in my book (and the specific suggestions I make in the second half of the book may apply equally well or badly here). In a way, the type-B materialist ought to be as concerned with the search for a “fundamental theory” as I am.

4 Epiphenomenalism and the paradox of phenomenal judgment

Hill & McLaughlin say that I endorse epiphenomenalism, and that my anti-materialist argument implies epiphenomenalism. This is not strictly true. In fact perhaps my preferred position on the mind-body problem (as Yablo in effect notes) is not epiphenomenalism but the “panprotopsychist” (or “Russellian”) position on which basic physical dispositions are grounded in basic phenomenal or protophenomenal properties. Far from making experience causally irrelevant, this view holds that experience will be part of the categorical grounds of causation.

I think there is also more to say about interactionism than I said in my book. In particular, I think there is no knockdown argument against a quantum interactionism on which consciousness is the categorical basis of wave-
function collapse. I don’t favor this view myself, but it needs to be taken seriously.

Both the panprotopsychist view and the quantum interactionist views are counterintuitive, but both are elegant and appealing and not obviously false. In both cases there are questions about whether they can really be made to work, but those issues remain open. So the conclusion of my anti-materialist argument ought not to be taken as epiphenomenalism. Rather, it is the disjunction of panprotopsychism, epiphenomenalism, and interactionism. The question between these three views is really a distinct issue, as it can only be decided by considerations further to the argument against materialism. For my part, I rank the options in roughly the order given, but it is a deep question on which I have no settled opinion.

On none of these positions is it true that we need “an independent swarming mass of fundamental psychophysical laws”, as Hill & McLaughlin suggest. We need fundamental psychophysical laws (or in the panprotopsychist case, we need principles associating physical dispositions with (proto)phenomenal bases), but there might be very few of them, perhaps only one. (Or two, in the interactionist case: one in each direction.) It is not unreasonable to suppose that there might be a single basic psychophysical principle (some sort of transformation operator, for example) that when applied to a specific complex physical system yields a specific complex phenomenal state.

It is true that I say that phenomenal states are in a sense explanatorily irrelevant to behavior; and I think that remains true on each of these views, as well as on the type-B materialist view. In each case, the very conceptual possibility of varying phenomenal states while preserving patterns of causal interaction suggests that explanations of behavior can be given in a way that makes no deep appeal to the phenomenal states in question. But that is not to say that they are causally irrelevant, except on the epiphenomenalist view.

I note also that Hill & McLaughlin are wrong in suggesting that I think sensory states are explanatorily irrelevant to our beliefs that we are in them. I do say that they are explanatorily irrelevant to the judgments we make about them, but this is in a stipulative functional sense of judgment, on which it comes to not much more than a disposition to report. I am quite clear, though, that experiences are not explanatorily irrelevant to our beliefs about them, and in fact often directly constitute aspects of those beliefs (see especially the last section of Chapter 5).

Shoemaker is right that my response to the paradox of phenomenal judgment is one of the weaker parts of the book. If I were writing Chapter 5 now, I would stress much more strongly the feature I have just mentioned: the role that phenomenal states play in constituting our phenomenal beliefs. While I noted this feature in the book, I tried to give an answer to the paradox that did not rely on it. I now think the constitutive relation between phenomenal
states and associated phenomenal beliefs is central to the epistemology here. (In effect, I would expand on the points in the last section of the chapter, particularly the parenthetical suggestion on pp. 207–8. See Chalmers 1998, lecture 3.) The paradox still applies to (functionally individuated) judgments and reports; but it is beliefs that are crucial in epistemology.

Shoemaker raises two very interesting issues related to the paradox. First, he suggests that a zombie’s phenomenal judgments will be causally related to physical states in a way that is sufficient for reference and knowledge; so a zombie’s judgments will refer to those states and will express knowledge of them; so my phenomenal judgments refer to physical states, too, as everything going on in the zombie is going on in me.

I deny the first step here. I think that for phenomenal concepts, no mere causal relation suffices for reference; and the same goes for a zombie’s analog of a phenomenal concept. As I say in the book, the causal theory of reference is by no means universal. In particular, the concepts to which it applies are generally causal concepts; at some level, their referents are conceived as the cause of something or other, if only implicitly. This is reflected by the fact that for most of our concepts (e.g. “atom” or “dog”), the notion that the referent of the concept is wholly epiphenomenal is epistemically impossible. It is epistemically impossible that all atoms or dogs are epiphenomenal. But phenomenal concepts are not causal concepts, as witnessed by the fact that epiphenomenalism about experience is epistemically possible, if counterintuitive. The same goes for a zombie’s analog of a phenomenal concept. So there is no reason to believe that a causal relation suffices for reference here, and in fact there is good reason to believe that a zombie’s phenomenal concepts do not refer. (In the zombie world, eliminativists are right!)

In his ingenious second point, Shoemaker notes that on my account, a zombie’s phenomenal judgments are false and unjustified, but the zombie does not seem to display any failure of rationality. But if the zombie’s judgment is both rational and unjustified, then the concept of justification is cut loose from its rational moorings. Thus a conceptual incoherence seems to threaten. This is a potential problem for any view on which zombies are conceptually possible, including not only my view but also those of Hill & McLaughlin, Loar, and Yablo. What should we say?

I think the best thing to say is that the zombie’s judgment is (a) unjustified (or at least not as justified as a conscious being’s corresponding judgment), but (b) at least negatively rational in the sense that the zombie makes no invalid inferences and other such errors of reasoning. Justification and negative rationality often go together, but not always; this is a case where justification and negative rationality come apart. (Other such cases might be provided by an externalist view of justification, where justification is not always grounded in rationality. This is not to say that the zombie case involves externalist justification: my analysis holds onto the internalist idea.
that justification is a matter of having accessible evidence.) I think it is plausible that where truly basic evidence is concerned, as with the relation between an experience and the associated phenomenal belief, the justification relation is something tighter than any cognitive relation involving mere negative rationality. This fits with the idea mentioned above, that the justification comes from the constitutive relation between experience and belief. If this is so, we should not expect the justification to be a wholly rational matter.

The zombie lacks evidence for its judgment (so it is unjustified), but it does not know that it lacks this evidence, and it is not culpable for this lack of knowledge (it is negatively rational). Shoemaker’s “rational moorings” for justification may still be preserved in a link between justification and positive rationality, which requires sufficient reasons and evidence for a judgment. But justification (especially basic justification) and positive rationality involve more than a mere absence of mistakes, so negative rationality does not in general suffice.

One could also take a more extreme line, and deny that notions such as rationality and justification apply to zombies at all. I have gradually become more sympathetic with the idea that phenomenology plays a role in constituting intentionality, so that a zombie has intentional states in at best a weak sense. But I do not want to rely on that point here.

5 Dancing qualia

Shoemaker also comments on the dancing qualia case. I think he is certainly right that qualitative beliefs will change when qualia change. This follows from the constitutive relation between experience and belief mentioned above; and it is how things must go, in order that the post-flip state not be incoherent. I did not mean to deny that beliefs would change in this sort of way. I was instead focusing on the absence of a belief about the change.

The surprising part of the dancing qualia case is not that the subject (after the flip) has false beliefs about past qualia, or even that he does not introspectively realize that his qualia are different. These things hold for any case of inverted qualia. The surprising part of the dancing qualia case is that the flip should go unnoticed when it happens. To have one’s qualia change massively before one’s eyes would be a major phenomenal event; but it would be a phenomenal event to which we have no cognitive access. Of course Shoemaker is right that if inverted qualia are possible, we should expect unnoticed changes in this scenario (because beliefs and memories will change too). But I don’t think this diminishes the implausibility of the consequent here; rather, it indirectly adds to the implausibility of the antecedent.

Shoemaker thinks the surprisingness can be explained away in this case, because there is an “unnatural tampering with cognitive dynamics”. I am not sure that it is all that unnatural. In future years, we might end up performing
this sort of substitution with some regularity; and on a lower level, physiological changes with little impact on function may already be happening all the time. It's true that if we assume that qualia are changing massively, then there is a sense in which the change is "unnatural", but the question of whether qualia change massively is precisely what is at issue. Characterized in neutral terms, the change to one's underlying processes does not seem especially unnatural; so it is surprising that it should produce such a massive, unnoticed change.

In any case the dancing qualia case was never intended as a knockdown argument, and there are certainly ways in which it can be resisted. On balance of intuition, I have come to think that in some ways the fading qualia argument is stronger than the dancing qualia argument: the idea of a system continually and completely out of touch with its "faded" qualia is even stranger than the idea of a system momentarily out of touch with its massively changing qualia. (Of course this is a conclusion Shoemaker would be happy with, as he thinks that absent and fading qualia are logically impossible, while inverted and dancing qualia may be naturally possible.) But at the end of the day these arguments are really intended as thought-experiments that any theory must face up to and make sense out of. Proponents of a different view might bite the bullet on any of them, as long as they are aware of the costs. I think that other things being equal, the conclusion that qualia do not fade and dance unnoticed is more plausible than the alternatives, but as with all of these matters, there is probably more to be said.\textsuperscript{12}

\textbf{References}


\textsuperscript{12} Thanks to Tamar Gendler, Alvin Goldman, Matthew Henken, David Lewis, Gregg Rosenberg, Daniel Stoljar, and Stephen Yablo for comments.