A Conceivability Argument

Adam Elga
Massachusetts Institute of Technology
Department of Linguistics and Philosophy
adam@mit.edu

November 8, 1998

1 Introduction

D. Chalmers has offered an argument against physicalism that proceeds from the conceivability of a *zombie world*, a physical duplicate of the actual world all of whose inhabitants have no phenomenal experiences. His argument depends on the *two-dimensional framework*, a semantic theory intended to systematically account for the divergence between the notions of a prioricity and necessity motivated by [8]. And it depends on a premise, CP, that connects conceivability to possibility.

I will briefly describe two-dimensionalism (§2), explain some difficulties in spelling out exactly which statement it is whose conceivability is supposed to refute physicalism (§3), object to CP (§4), and respond to Chalmers's defense of CP against the sort of objection I describe (§5).

2 Two-dimensionalism

This section gives a brief gloss of two-dimensionalism (mostly to get terminology straight). For a fuller treatment, see [3, pp. 56–64].

2.1 Concepts

One might think that understanding a concept such as 'water' requires grasping its dependency function: the function that associates to each possible world the extension of the concept in that world (at least for sufficiently sophisticated agents who possess the concept 'intension'—pace Soames [9]). According to two-dimensionalism, however, the dependency function associated with a concept typically depends on the context in which a term denoting the concept is uttered. Since it is too strict to require that hearers know what the context of utterance is in order to understand a term, two-dimensionalists substitute the following more modest requirement: understanding a term (of language L) requires grasping the way that the dependency function of the concept associated with the term depends on the term's context of utterance.

On this way of doing things, determining the extension of a concept requires specifying both a context of utterance and a world of evaluation. The *2D intension* of a concept is a two place function $f(\cdot,\cdot)$ intended to capture this dependence. For a context c and world W, $f(c, W)$ is the extension at W of a term denoting the concept (still considered as a word of language L) as uttered in c. Although a context of utterance is ordinarily taken to be a world along with a specification of a privileged individual, for simplicity I shall take a context to simply be a possible world.

If f is the 2D-intension of a concept T, then the *primary intension* of T is the function that maps each context c to $f(c, c)$. The *secondary intension* of T is the function that maps each world W to $f(@, W)$ where @ is the actual context.

2.2 Statements

The 2D-intensions of concepts employed in a statement combine compositionally to determine the 2D-intension of the statement, a two place function that associates with each context of utterance and world of evaluation the truth value of the statement as uttered in that context and evaluated at that world. [11] If g is the 2D-intension of a statement S, then the *primary intension* of S is the function that maps each context c to $g(c, c)$. The *secondary intension* of S is the function that maps each world W to $f(@, W)$ where @ is the actual context.

Some further terminology: To *take a world as actual* is to treat it as the context of utterance. To *take a world as counterfactual* is to treat it as the world of evaluation.[1] A world W is said to *verify* a statement S if the

[1] [1, p. 26] attributes this terminology to [6].

primary intension of S assigns True to W. And W is said to *satisfy* S if the secondary intension of S assigns True to W.[2]

The two-dimensional apparatus provides a neat explanation of how even an extremely keen logician might hear a necessary statement, understand the statement, and yet not thereby learn anything that would allow her to deductively infer that the statement expresses a truth. The explanation is this: in understanding the statement, what the logician is required to grasp is the 2D-intension g of the statement. The statement is necessary in the sense that its *secondary* intension is necessary—i.e. $g(@, \cdot)$ assigns True to every possible world—but the logician may not know which context is actual. If for all she knows the context is c', and if the primary intension of S assigns False to c', then without doing empirical research to rule out the hypothesis that the actual context is c', she has no way of inferring that S expresses a truth. In this case S has a necessary secondary intension but is *a posteriori*—mere understanding of S (without empirical fieldwork) does not provide anyone with sufficient knowledge to deductively infer that S expresses a truth.

To summarize: the two dimensional framework allows for a posteriori necessities in at least this case: a statement has a necessary secondary intension but has a contingent primary intension. Is this the *only* way that a posteriori necessities can arise? Sometimes "two dimensionalism" is taken to include the claim that it is. In that case any statement whose primary intension is necessary would thereby qualify as being a priori[3]. But I shall take the commitments one accepts when accepting the two dimensional framework to be more modest. In accepting the two dimensional framework, one is required to agree that *one* way for a statement whose secondary intension is necessary to be a posteriori is for its primary intension to be contingent. But it is compatible with the two dimensional framework as I shall understand it to believe something that Chalmers denies—that some statements with necessary primary intensions are a posteriori.

[2]This terminology is from [11].

[3](i.e., not a posteriori, as explicated above)

3 What is the zombie statement?

3.1 Physicalism

The target of Chalmers's argument is *physicalism*, understood as follows. Call the stock of concepts featured in some not-too-different but appropriately improved version of present-day physics the *fundamental physical* ones. At a first approximation, physicalism is the claim that a complete specification of the pattern of instantiation of the fundamental physical concepts in the actual world is a full specification of the actual world. This analysis needs fine-tuning, but for present purposes we needn't bother. What matters is that physicalism is incompatible with the existence of a possible world that is a physical duplicate of the actual world but which contains no pain.

Chalmers's argument depends on the conceivability of a statement asserting that such a scenario obtains. But which statement is this, exactly?

3.2 Proposal A

Let the expression "$\cdots\phi_1\cdots\phi_2\cdots\phi_3\cdots$" abbreviate a complete physical specification of the actual world, where the ϕs are terms denoting fundamental physical concepts. Then our first zombie statement is[4]

Z $\cdots\phi_1\cdots\phi_2\cdots\phi_3\cdots$ and there is no pain.

Consider Argument A:

A1 If a statement S is conceivable, then some possible world verifies S.

A2 Z is conceivable.

A3 'pain' has identical primary and secondary intensions.[5]

A4 Every fundamental physical concept has identical primary and secondary intensions.

[4] Ordinarily, and throughout most of this paper, the label to the left of a displayed sentence names the proposition the sentence expresses. The sole exception is for zombie statements: in this paper a label that consist of the letter 'Z' (perhaps with decorations) names the *statement* associated with the sentence to the right of the label.

[5] This premise is stronger than it need be, but I will grant the stronger version here and elsewhere to simplify the discussion.

A5 There is a possible world W_Z that verifies Z. (A1, A2)

A6 Physicalism is false. (A3, A4, A5)

Argument A is valid.[6] A3 and A4 guarantee that Z has identical primary and secondary intensions, which allows us to conclude from A5 that W_Z satisfies Z, which makes W_Z a counterexample to physicalism.

Should a physicalist seek to avoid Argument A's conclusion by denying A4, she will be forced to adopt what Chalmers calls *panprotopsychism.* The panprotopsychist holds that physical concepts have distinct primary and secondary intensions. This allows her to block the inference from A5 to A6—just because a world verifies Z doesn't mean that it satisfies Z. Pretending for the moment that 'electron' is a fundamental physical concept, here is how a Panprotopsychist's motivating story might go:[7]

In any world, the concept 'electron' picks out the objects with the same intrinsic character as the ones that fill the electron-role in the actual world. In our world electrons have an intrinsic property ψ. ψ is hidden in the sense that the presence or absence of it in an electron is entirely independent of the physical causal interactions between electrons and other particles.

Now consider the world W_Z delivered by A5, above. W_Z verifies Z, so it agrees with the actual world on physical concepts, if we evaluate physical concepts by primary intension. But this is not enough to guarantee, for example, that the electrons in W_Z have ψ. W_Z need not agree with the actual world on physical concepts, if we evaluate physical concepts by secondary intension. So granting the existence of W_Z leaves available a version of physicalism—one that rules out the existence of pain-free physical duplicates of the actual world *where "physical duplicate" means "agreeing on all physical concepts, evaluated according to secondary intension"*. On this story, W_Z is pain-free, but it is not a physical duplicate of the actual world. And what explains both of these facts is that the electrons in W_Z don't have ψ. [3, p. 135] (Because their absence alone is enough to make the difference between a world with pain and a world without it, Chalmers calls the intrinsic properties postulated by the panprotopsychist "protophenomenal properties".)

Although panprotopsychism is consistent with a version of physicalism, Chalmers claims, and let us grant, that allowing that phenomenal properties

---

[6]—modulo the assumption that the actual world contains pain, which I shall take for granted throughout.

[7]Here I follow [3, pp. 135–6].

depend on the hidden intrinsic properties of, say, fundamental particles, is decidedly against the spirit of physicalism.

So: Argument A is valid, and a physicalist can't comfortably avoid its conclusion by denying A4, on pain of being forced into panprotopsychism. Let us postpone doubts about A1 until §4, and take A3 for granted. What's left is A2, the premise that asserts that our zombie statement Z is conceivable. Some comments about Z:

1. We can't write Z down explicitly because the word is not yet in on which are the fundamental physical concepts.

2. Even if we did know the list of fundamental physical concepts, we still wouldn't be able to write down Z since it is almost certainly horrendously long. If "suitably idealized physics" is anything like present physics, Z is infinitely long.

3. Z would be extremely difficult to determine, involving as it does the measurement of the detailed physical state of the universe from the beginning to the end of time.

For these reasons, no human has ever or will ever grasp Z and thus no human has conceived that Z is true. A2 is thus extremely implausible.[8]

3.3 Proposal B

Let us construct an argument that depends on a shorter zombie statement—one that we humans can explicitly grasp. Here is a candidate:

[8] Alex Byrne has suggested to me that anyone who is convinced by Argument A ought to regard themselves in a position to refute panprotopsychism with a similar concevability argument—one that substitutes for the zombie statement Z the statement

Z_ψ $\quad \cdots \phi_1 \cdots \phi_2 \cdots \phi_3 \cdots \psi_1 \cdots \psi_2 \cdots$ and there is no pain.

(Here the expression "$\cdots \psi_1 \cdots \psi_2 \cdots$" abbreviates a complete specification of the pattern of instantiation of the hidden intrinsic properties postulated by the panprotopsychist.)

The conceivability argument aimed at refuting the panprotopsychist depends on a premise that asserts the conceivability of Z_ψ. But even if one thought the difficulties in grasping Z described above could somehow be overcome, one might still think that there are additional difficulties in grasping Z_ψ. One might think that it is impossible for us to manage to refer to specific protophenomenal properties since they are outside the domain of physical investigation.

Z′ The world has exactly the same fundamental physical properties as the actual world, and there is no pain.

Let us replace Z by Z′ in premises A2 and A5 to produce Argument B:

B1 If a statement S is conceivable, then some possible world verifies S.

B2 Z′ is conceivable.

B3 'pain' has identical primary and secondary intensions.

B4 Every fundamental physical concept has identical primary and secondary intensions.

B5 There is a possible world $W_{Z'}$ that verifies Z′. (B1, B2)

B6 Physicalism is false. (B3, B4, B5)?

Since we can explicitly write down Z′, and since Z′ is so much shorter than Z, B2 is much more plausible than A2. However, Argument B is invalid since the truth of physicalism is compatible with B3, B4, and B5. To see this, recall how Argument A works. A3 and A4 guarantee that Z has identical primary and secondary intensions. So when A5 delivers a world W_Z that verifies Z, we are able to infer that W_Z satisfies Z, which makes W_Z a counterexample to physicalism. In contrast, B3 and B4 do *not* guarantee that Z′ has identical primary and secondary intensions. (They had better not, because Z′ clearly has distinct primary and secondary intensions.) The antecedent of Z′ has a necessary primary intension: for any world W, taking W as actual makes "The world has exactly the same fundamental physical properties as the actual world" true at W, since every world is a physical duplicate of itself. So any world in which there is no pain verifies Z′. Since the existence of pain-free worlds is consistent with physicalism, B6 does not follow from B3, B4, and B5.

3.4 Proposal C

Next try: let us introduce the term "$W_@$" to have a 2D-intension that is a constant function. Taking any context as actual and any world as counterfactual, "$W_@$" denotes the world that you, the reader, inhabit. The following matrix[9] represents (part of) the intension of "$W_@$".[10]

	W_1	W_2	W_3
W_1	@	@	@
W_2	@	@	@
W_3	@	@	@

We can use "$W_@$" to form a new zombie statement:

Z″ The world has exactly the same fundamental physical properties as $W_@$ and there is no pain.

and Argument C:

C1 If a statement S is conceivable, then some possible world verifies S.

C2 Z″ is conceivable.

C3 'pain' has identical primary and secondary intensions.

C4 Every fundamental physical concept has identical primary and secondary intensions.

C5 There is a possible world $W_{Z''}$ that verifies Z″. (C1, C2)

C6 Physicalism is false. (C3, C4, C5)?

But Argument C is still invalid. The trouble is that C3 and C4 leave it open that the concept 'fundamental physical property' has distinct primary and secondary intensions, and hence leave it open that Z″ has distinct primary and secondary intensions. Suppose we replace C4 by

C4′ The concept 'fundamental physical property' has identical primary and secondary intensions.

---

[9] The entry of the matrix in row i, column j is to be interpreted as the extension of "$W_@$" at W_j relative to context W_i.

[10] The existence of a term such as "$W_@$" might be thought to go against the spirit of two-dimensionalism and I don't mean to suggest that Chalmers would endorse the use of such a term. It's just that I don't see how to form an appropriate zombie statement without the use of such a term.

The resulting argument (call it Argument C$'$) is valid. However, a physicalist may avoid its conclusion by denying C4$'$. Denying C4$'$ is quite reasonable, and doesn't require a committment to anything like panprotopsychism. Here is a story one might tell to motivate the denial of C4$'$ to block Argument C$'$:

In any world, the concept 'fundamental physical property' has as its extension the properties that figure in the suitably idealized physics of the actual world.

Let $W_\emptyset$ be an empty world (which contains no contingently existing entities). The best physics of $W_\emptyset$ has nothing to say at all, and so no properties are in the primary intension of 'fundamental physical property' evaluated at $W_\emptyset$. Therefore $W_\emptyset$ trivially verifies Z$''$—taking $W_\emptyset$ as actual, all worlds are physical duplicates since there are no fundamental physical properties. But $W_\emptyset$ is no threat to physicalism.

3.5 Proposal D

Final try: Introduce "$\Phi_@$", another term whose 2D-intension is a constant function. Stipulate that taking any world as actual and any world as counterfactual, "$\Phi_@$" has as its extension the properties that count as physical in the world that you, the reader, inhabit. Now we can construct

Z$'''$ The world is a duplicate of $W_@$ as far as $\Phi_@$-properties go and there is no pain.

and use it in Argument D:

D1 If a statement S is conceivable, then some possible world verifies S.

D2 Z$'''$ is conceivable.

D3 'pain' has identical primary and secondary intensions.

D5 There is a possible world $W_{Z'''}$ that verifies Z$'''$ (D1, D2)

D6 Physicalism is false. (D3, D5)

Argument D is valid. The next section will examine the joint plausibility of D1 and D2.

4 A criticism of CP

Chalmers connects conceivability and possibility with the following premise:

CP If a statement S is conceivable, then some possible world verifies S.

The force of CP depends upon the notion of conceivability that figures in it. Chalmers introduces conceivability as follows:

> Let us say that a statement is conceivable (or conceivably true) if it is true in some conceivable world.... On this view of conceivability, the conceivability of a statement involves two things: first, the conceivability of a relevant world, and second, the truth of the statements in that world. [3, pp. 66-7]

Let us call the notion of conceivability so introduced "strong conceivability". First note: according to Chalmers, a "conceivable world" is a type of possible world—a way things might possibly have been. So strong conceivability is a success term in the following sense: acts of strong conceiving are by definition mental acts whose objects are *some possibility or other*. If S is impossible, then it is built into the definition of strong conceivability that no one can manage to conceive S.

Second note: connecting this up with the two-dimensional framework means distinguishing two notions of strong conceivability "which we might call *1-conceivability* and *2-conceivability*, depending on whether we evaluate a statement in a conceivable world according to the primary or secondary [intension of the statement]". [3, p. 66] Everything in the present paper concerns 1-conceivability, so throughout I leave the "1-" modifier implicit.

Considerations of what is strongly conceivable are not the sort of considerations that help settle debates over disputed claims of possibility. Rather, they are part of what is disputed.

Consider the plausibility of versions of CP and ZC that use strong conceivability:

CP_s If a statement S is strongly conceivable, then some possible world verifies S.

ZC_s The zombie statement[11] is strongly conceivable.

[11] Here and from now on I will ignore any worries about exactly which statement the zombie statement is. I will suppose for the sake of argument that some appropriate zombie statement has been identified.

CP_s is true by definition—no problems there. But since the possibility of the zombie statement is in dispute, ZC_s is difficult to support. Chalmers motivates the conceivability of the zombie statement as follows: "it certainly seems that a coherent situation is described; I can discern no contradiction in that description". But this sort of consideration is insufficient without another premise that connects the apparent coherence of a situation with its strong conceivability.[12]

Rather than introduce such a premise, let us use a weaker notion of conceivability. Say that a statement is *weakly conceivable* if an agent's full understanding of S is not sufficient for her to deductively infer that the primary intension of S is impossible, even if her deductive powers are greatly idealized.[13] Then we have

CP_w If a statement S is weakly conceivable, then some possible world verifies S.

ZC_w The zombie statement is weakly conceivable.

CP_w has some initial plausibility. Remember that according to the two-dimensional framework, understanding a statement requires grasping its 2D intension and hence grasping its primary truth condition. Now consider the contrapositive of CP_w: if no possible world verifies S, then S is not weakly

---

[12]The same point applies to any analysis of conceivability according to which it is true by definition that if a statement S is conceivable then there is a possible world that verifies S. For example, in [4, §1.5], Chalmers distinguishes between the *positive conceivability* of S ("clear and distinct conceivability of a situation verifying S") and the *negative conceivability* of S ("absence of any apparent contradiction in S"). If positive conceivability is a success term, then the premise

ZC_p The zombie statement is positively conceivable.

is difficult to support. In conceiving a statement S, I do not see what mental act I can perform over and above (i) imagining a situation that I take to verify S and (ii) trying hard and failing to find either something contradictory in that situation or something that would stop that situation from being embedded in a possible world. If when I conceive the zombie statement all that I accomplish is (i) and (ii) then my act of conceiving doesn't ensure that S is positively conceivable, in which case I am entitled to reject ZC_p.

On the other hand, if positive conceivability is *not* a success term, then the following discussion (about versions of ZC and CP that use weak conceivability) also applies to versions of ZC and CP that use positive conceivability. (I am indebted to Stephen Yablo for discussion on this point.)

[13]Cf. the Laplacian demon featured in [1].

conceivable. Suppose that no possible world verifies S—i.e. suppose that S's primary intension is impossible. We just said that understanding S requires grasping its primary intension. So it seems that understanding S requires understanding that S's primary intension is impossible.

To see where this reasoning goes wrong,[14] distinguish between two readings of the claim that understanding a statement requires grasping its 2D-intension. To make the relevant contrast, ignore the two-dimensional framework for the moment and just consider what it takes to grasp (ordinary) truth-conditions. Suppose I say to you "This room is full of flurgs," although you only know a few facts about flurgs. (You know, for example, that flurgs are animals small enough to fit in a shoe-box, but don't know how many toes they have, etc.) In a certain sense (call it the *ordinary sense*) you nevertheless grasp the truth-conditions of the proposition I've expressed: for starters, you know that it is true in those worlds where the room is full of flurgs. On the other hand, this sort of understanding does not guarantee that, given a set of worlds picked out by a description (in *any* vocabulary that you understand), you are able to tell whether that set of worlds is the set of worlds in which what I've said is true.

Alternatively, one might say that to grasp the intension of a proposition is to have the ability to tell of a set of worlds (again, picked out by a description in any vocabulary that you understand) whether that set is the set of worlds in which the proposition is true. Call this the *extraordinary sense* of "grasping truth-conditions".

Consider two-dimensionalism's requirement that an understander of S grasp S's primary intension. If the grasping in question is mere ordinary grasping, then the motivation for CP_s above fails. As an example, suppose that the primary intension of a statement T is impossible. Suppose further that T is the conjunction of statements C and D, and that C and D both have contingent primary intension.

In order to grasp the primary intension of T, it may be sufficient to

1. understand what it takes for a world to verify C;
2. understand what it takes for a world to verify D;
3. know the truth table for "and",

[14] The criticism of CP described in the remainder of this section is based on [11].

where the sort of understanding required by (1) and (2) is something less than grasping in the extraordinary sense.

An agent may only be able to tell whether a world verifies C when it is presented in thought to her under one sort of description, and may only be able to tell whether a world verifies D when it is presented under a different sort of description. Suppose that world W verifies C but not D. Then, under one sort of description, the agent may be able to see that W verifies C, and under a different sort of description may be able to see that W doesn't verify D. But since W is presented to her under two different sorts of descriptions, she may not be aware that it is the very same world that has been presented to her twice, and thus may not be able to deduce that W doesn't verify T, even though it in fact doesn't.

Such an agent grasps the primary intension of T in the ordinary sense, but is not thereby able to deduce that the primary truth condition of T is impossible.[15]

On the other hand, if two-dimensionalism demands that understanding a statement requires one to grasp its primary intension in the extraordinary sense, then it becomes doubtful whether any of us has managed to grasp the zombie statement. We do imagine worlds under descriptions and are not always able to reidentify worlds that have been presented to us under two disparate sorts of descriptions. In this case we should not accept ZC_w.

5 Defending CP_w

Denying CP_w means admitting that there are what Chalmers calls *strong necessities*: weakly conceivable statements whose primary intensions are impossible. In response to a criticism of CP_w along the lines of the one described in the previous section ([11]), Chalmers has argued against strong necessities [2].

In order to explain Chalmers's argument against strong necessities, suppose that we provisionally endorse the existence of strong necessities. On our view the zombie statement is weakly conceivable, but has an impossible primary intension. We grant that there are two varieties of possibility of *statements*,[16] but insists that when it comes to *worlds*, there is only one sort

[15] This example is based on one Yablo gives in [11].

[16] Namely: possibility according to primary and secondary intension respectively.

of possibility: possible worlds are all ways the world might have been. And we start out holding that no possible world verifies the zombie statement.

Chalmers thinks such a view will get us into trouble because

> 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 inference involving consciousness, of the contents of rational beliefs about consciousness, and so on... talk of logically possible zombie worlds is justified in the usual way by their role in these uses. [2, §3.2]

This passage suggests a route by which Chalmers thinks we will be forced to expand our domain of possible worlds to include (for example) a possible world that verifies the zombie statement. After so expanding our domain of possible worlds, we might try to keep a distinction between the worlds we originally thought possible and those we were forced to add later by calling the former worlds "*metaphysically* possible". But to do that would be to introduce a new grade of possiblity of *worlds* (not statements), and thereby to endorse an implausible "modal dualism" [2, §3.5].

Since I find Chalmers's arguments against modal dualism utterly convincing, I concede that if we were required to expand our domain of possible worlds in the way Chalmers suggests, then that would show that our initial endorsment of strong necessities was unacceptable.

The route by which Chalmers thinks we are required to expand our domain of possible worlds divides into two stages:

Stage 1 We start out with the view that there is just one sort of possibility of worlds, and that none of the worlds possible in this one sense verify the zombie statement. *We find that our existing setup is inadequate to model the contents of beliefs about consciousness, among other phenomena.* In order to remedy this, we are forced to add to our stock of primitives new world-like entities[17] whose job it is to verify statements that are weakly conceivable but are not verified by any possible world.

[17] I use the term "world-like entity" rather than "world" or "possible world" in order to avoid prejudging the issue of whether these entities represent ways the world might have been.

Stage 2 We are able to use our new expanded set D of world-like entities (our original domain of possible worlds plus the new world-like entities we added in Stage 1) to model the contents of beliefs about consciousness and to play the role of the domain of worlds in our analyses of counterfactuals and "rational inference". *But "deep constitutive connections between modality and rationality" [2, §3.5] require that we count all the members of D as representing genuine possibilities, in virtue of the theoretical role D plays in the modeling and analyses just mentioned.*

Each of these stages depends on a crucial claim, which I have italicized. I will argue that the both crucial claims are false, and hence that Chalmers's argument fails to show that there is anything wrong with believing in strong necessities. §5.1 argues against the crucial claim of Stage 1. §5.2 grants the crucial claim of Stage 1 and argues against the crucial claim of Stage 2.

5.1 The crucial claim of Stage 1

We start out with the view that there is just one sort of possibility when it comes to worlds, and that none of the worlds possible in this one sense verify the zombie statement. Chalmers contends that we must add new world-like entities to our domain of possible worlds in order to make available to ourselves reasonable analyses of the contents of belief, counterfactuals, and rational inference. The task of the remainder of this section is to show that we can accomplish these tasks without adding new world-like entities to our domain of possible worlds.

5.1.1 Belief states

We are faced with Joe, who believes that the actual world is a zombie world. In modeling Joe's beliefs we need not add to our domain of possible worlds. Here are two reasons:

First, in using possible worlds as part of our semantic theory, we are not thereby committed to identifying belief states with sets of possible worlds. We might, for example, model Joe's belief state as pair of objects whose first member is a set of possible worlds and whose second member is some device used to distinguish between belief states with identical intensions. On this way of doing things, the fact that Joe believes an impossibility would just mean that the pair we use to model his belief state has the empty set as its first member. Or we might use some other method entirely—one that does not use sets of possible worlds at all. For example, we might model his belief state as a set of sentences in some language, in which case it might include the sentence "the actual world verifies the zombie statement". This is entirely compatible with talking as though belief-states were sets of possible worlds *in some theoretical contexts*: namely, those contexts in which we don't care about the differences between belief-states that have the same truth-conditions.

Second, even if the (widely rejected) practice of identifying belief states with sets of possible worlds were mandatory, we still might be able to model Joe's belief adequately without expanding our domain of worlds. True, we would have some work to do, but we would have this work to do anyway in modeling other necessary or impossible beliefs, most notably mathematical beliefs. We might, for example, attempt to reconstrue Joe's belief as the belief that the sentence he uses to make the zombie statement expresses a truth. (See [10] for a treatment of this and other strategies to model non-contingent beliefs under identification of belief states with sets of possible worlds.)

So much for belief states. How about counterfactuals—would we be able to offer a reasonable analysis of them?

5.1.2 Counterfactuals

Consider first the class of counterfactuals whose antecedents we initially regard as having impossible primary intensions (e.g., the counterfactual whose antecedent is the zombie statement and whose consequent is "stubbing one's toes doesn't cause pain"). Chalmers challenges us to give an analysis of these without expanding our domain of possible worlds. Two analogous points hold:

First, it is not mandatory that we use some (Stalnaker or Lewis-style) analysis of counterfactuals in terms of possible worlds to analyze *all* counterfactuals. The counterfactuals in question have antecedents that we regard as having impossible primary intensions—this is reasonable grounds for giving them a different style of analysis. Compare: friends of possible-worlds analyses of counterfactuals are not obliged to apply those analyses to counterfactuals whose antecedents are explicit contradictions. One available position is to endorse a possible-worlds analysis for counterfactuals with possible antecedents, but to endorse some other analysis for counterfactuals whose antecedents are impossible.

Second, even if it were mandatory to use a possible-worlds analysis for all counterfactuals, we still would not be forced to expand our domain of possible worlds. Suppose that our possible-worlds analysis of counterfactuals implies that every counterfactual with an impossible antecedent is true. Then we would disagree (on the truth values of counterfactuals with zombie-statement antecedents) with people who think that some possible world verifies the zombie statement. But this is as it should be—counterfactuals are modal statements, and so it is natural that disagreement about what is possible leads to disagreement about which counterfactuals are true.

The same goes for counterfactuals whose antecedents we initially regard as contingent. Someone who thinks there is a possible world W_Z that verifies the zombie statement might think that the counterfactual "If there had been no pain in the world, then there still would have been humans" is true because W_Z is the pain-free world most similar to the actual world. A type-B materialist might regard that counterfactual as false on the grounds that the pain-free world most similar to the actual world contains no humans. Again, modal disagreement leads to disagreement about the truth-values of counterfactuals. No problem.

5.1.3 Rational inference

Chalmers claims that a defender of strong necessities would need to count additional worlds as possible in order to "make sense of rational inferences involving consciousness". I think this is put misleadingly. Chalmers is not here interested in questions about what it is rational to infer from what. For example, it is often rational to infer statements that are not entailed by statements that one believes, and it is often irrational to infer statements that are entailed by what one believes ([7]). Rational inference is beside the point; what we should understand Chalmers as demanding of type-B materialists is an account of the relations of a priori entailment that hold among statements having to do with consciousness.

Chalmers does not spell out why he thinks these a priori entailment relations are difficult for a type-B materialist to account for, but his remarks suggest an argument like this one:

The most straightforward way to analyze a priori entailment within the two-dimensional framework is

EN For any statements R and S, R a priori entails S iff every world that verifies R also verifies S.

If a type-B materialist were to accept this analysis, she would conclude that the zombie statement a priori entails every statement (since by her lights, no world verifies the zombie statement). But Chalmers thinks this conclusion is implausible because "we can write coherent science fiction about zombies, and speak coherently about the truth in such fictions" and because "in zombie worlds, statements can be semantically evaluated with ease, and they never come out both true and false" ([2]).

However, anyone who denies CP_w on the basis of the sort of opacity phenomena described in §4 ought to reject EN in favor of the following competing analysis:

EN′ For any statements R and S, R a priori entails S iff $\ulcorner R \supset S \urcorner$ is a priori.

On the assumption that CP_w is true, EN and EN′ come to the same thing. But if CP_w fails then the two analyses diverge. We concluded in §4 that the problem of reidentifying worlds under different sorts of presentations allows for strong necessities. We should likewise conclude that the same problem allows for a pair of statements R, S such that every world that verifies R verifies S, but such that an agent's understanding of R and S does not enable her to deduce that every world that verifies R verifies S. Adopting EN′ on the basis of this observation, we avoid the unpleasant conclusion that the zombie statement a priori entails every statement.

5.2 Stage 2

The previous section argued that we (as believers in strong necessities who take it that no possible world verifies the zombie statement) can model belief states, offer an analysis of counterfactuals, and offer an analysis of a priori entailment—all without adding to our domain of possible worlds.

For the sake of argument suppose that we cannot. Suppose that in order to accomplish the theoretical tasks just mentioned we are forced to add to our domain of worlds new world-like entities. The new entities have the job of verifying those weakly conceivable statements not verified by any members of our original domain of worlds. The question is: after adding to our domain of worlds, are we thereby committed to counting the entities added as describing ways the world might have been? I will argue that we are not so committed.

Consider first how we use our domain of worlds in modeling belief states. Recall Joe, who believes the zombie statement. If we model belief states as sets of world-like entities from our domain, then we will model Joe's belief-state as a nonempty set of world-like entities. But doing this does not commit us as theorists to the claim that every element of that set describes a way the world might have been. We might reasonably continue to regard Joe as believing something impossible.[18] If we do so, we will regard the world-like entities we use to model Joe's belief state as merely that part of our theoretical apparatus useful for representing the differences between believing one impossibility and another.

Next suppose that we endorse a Stalnaker- or Lewis- style analysis of counterfactuals with an associated domain that contains, for example, a world-like entity W_Z added in order to verify the zombie statement. Then our analysis would evaluate as false the counterfactual whose antecedent is the zombie statement and whose consequent is "humans do not exist". And it would evaluate as true the counterfactual whose antecedent is the zombie statement and whose consequent is "stubbing one's toe doesn't hurt". But we might reasonably continue to regard the zombie statement as impossible. If we do, W_Z will earn its keep in our theoretical apparatus not by describing a way the world might have been, but by aiding in the evaluation of certain counterfactuals with impossible antecedents.

Chalmers offers the following consideration against the sort of proposals just described:

> But even on [an extreme modal-realist] 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 counterfactual worlds are regarded as ersatz abstract objects, there seems to be no grounds for resistance. It is easy to construct an ersatz object that behaves in just the way the zombie world should. The obvious strategy is to use maximal consistent worldbooks, where "consistent" is understood in the a priori sense. [2, §3.2]

Assume that we subscribe to the dominant view that possible worlds are abstract objects. Then the scenario under consideration is this: we start

[18] Here and in the remainder of this section I shall mean by "impossible" impossible-according-to-primary-intension.

out believing that no possible world verifies the zombie statement. Next we are forced to add to our domain of worlds new world-like objects that verify statements such as the zombie statement. Chalmers notes that these new world-like entities will be qualitatively just like the the possible worlds we started with: the worlds in our original domain are sets of interpreted sentences (say) and so are the new world-like entities. He says that "there seems to be no ground for resistance". But resistance to what? We certainly have no ground for denying that abstract entities like the ones we added exist. (If we accept the existence of some sets of interpreted sentences it would indeed seem arbitrary to deny the existence of other sets of interpreted sentences.) But that is irrelevant. The relevant issue is whether we can reasonably resist the claim that all of the new world-like entities represent ways the world might have been. And we can reasonably resist this claim.

We can resist by drawing attention again to the opacity phenomenon (described in §4) that motivated us to deny CP_w in the first place. Consider again the example of a statement T (which has an impossible primary intension) that is the conjunction of statements C and D, each of which has a contingent primary intension. No possible world verifies T, and yet someone who grasps T's primary intension (in the ordinary sense) might be unable to deduce that this is so. She might be unable to deduce this because it can happen that after she conceives of a world W once under one description and conceives of what is in fact W again under a different description, she has no way of figuring out that she has conceived of the same world twice.

Suppose now that we are forced to add to our domain a new world-like entity W_T whose job it is to verify T. Chalmers notes that if we were to start regarding W_T as representing a way the world might have been, then we wouldn't have the sorts of technical troubles we would have if (for example) we started regarding as possible a world in which 2+2=5.[19] Even if this is right, the fact that no technical obstacle *prevents* us from regarding W_T as representing a possibility is no positive reason to *start* regarding W_T as representing a possibility.

Available to us is a reasonable explanation of how a statement not verified by any possible world might nevertheless be weakly conceivable (namely: the

[19] He writes: "I suppose someone 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. [2]"

opacity explanation just given). Given that we are expanding our domain of worlds to include enough world-like entities to verify every weakly conceivable statement, that same explanation gives us good reason to think that not all of those new entities describe ways the world might have been.

It might be objected that we have ended up endorsing a kind of modal dualism. Our domain does indedd contain world-like entities of two types: (i) our original domain of worlds, and (ii) the world-like entities we were forced to add. Aren't we then comitted to two grades of possibility? The answer is that we are not. We still think that there is just one sort of possibility when it comes to worlds, and that exactly those worlds in our original domain possess it. The scenarios described by the new world-like entities we added to our domain are just plain impossible—as impossible, if you like, as scenarios in which 2+2=5. The difference between the zombie scenario and a scenario in which 2+2=5 has nothing to do with degrees of possibility. Instead the difference is that our understanding of '2+2=5' allows us to immediately deduce that it has an impossible primary intension, while our understanding of the zombie statement leaves it open that its primary intension is possible.

References

[1] Alex Byrne. Cosmic hermeneutics. Manuscript, 1998.

[2] David J. Chalmers. Materialism and the metaphysics of modality. Reply to participants in symposium on The Conscious Mind.

[3] David J. Chalmers. *The Conscious Mind.* Oxford Univ. Press, 1996.

[4] David J. Chalmers. Mind and modality. Lectures given at Princeton University, 1998.

[5] David J. Chalmers. Modal rationalism and the mind-body problem. Unpublished notes for address at Notre Dame, 1998.

[6] M. K. Davies and I. L. Humberstone. Two notions of necessity. *Phil. Stud.*, 38:1–30, 1980.

[7] Gilbert Harman. *Change in view: principles of reasoning.* MIT Press, Cambridge, Mass., 1986.

[8] Saul A. Kripke. *Naming and Necessity.* Harvard Univ. Press, 1972.

[9] Scott Soames. Truth and understanding. *Philosophical studies*, (65):17–35, 1992.

[10] Robert Stalnaker. *Inquiry.* MIT Press, Cambridge, Mass., 1984.

[11] Stephen Yablo. Concepts and consciousness
. http://web.mit.edu/yablo/www/concon.html, 1998.