The Two-Dimensional Argument Against Materialism

David J. Chalmers

A number of popular arguments for dualism start from a premise about an epistemic gap between physical truths about truths about consciousness, and infer an ontological gap between physical processes and consciousness. Arguments of this sort include the conceivability argument, the knowledge argument, the explanatory-gap argument, and the property dualism argument. Such arguments are often resisted on the grounds that epistemic premises do not entail ontological conclusion.

My view is that one can legitimately infer ontological conclusions from epistemic premises, if one is careful about how one reasons. To do so, the best way is to reason first from epistemic premises to modal conclusions (about necessity and possibility), and from there to ontological conclusions. Here, the crucial issue is the link between the epistemic and modal domains. How can one reason from theses about what is knowable or conceivable to theses about what is necessary or possible?

To bridge the epistemic and modal domains, the framework of two-dimensional semantics can play a central role. I have used this framework in earlier work (Chalmers 1996) to mount an argument against materialism. Here, I want to revisit the argument, laying it out in a more explicit and careful form, and responding to a number of objections. In what follows I will concentrate mostly on the conceivability argument. I think that very similar considerations apply to the other arguments mentioned above, however. In the final section of the paper, I show how this analysis might yield a unified treatment of a number of anti-materialist arguments.

1 The Conceivability Argument

The most straightforward form of the conceivability argument against materialism runs as follows.

[0]An abridged version of this paper is forthcoming in B. McLaughlin (ed.) *The Oxford Handbook of the Philosophy of Mind*. The full version is forthcoming in my *The Character of Consciousness* (Oxford University Press). Some material in this paper is drawn from Chalmers 1999, 2002, 2004b, and 2005.

(1) $P\&\neg Q$ is conceivable

(2) If $P\&\neg Q$ is conceivable, $P\&\neg Q$ is metaphysically possible

(3) If $P\&\neg Q$ is metaphysically possible, materialism is false.

(4) Materialism is false.

Here P is the conjunction of all microphysical truths about the universe, specifying the fundamental features of every fundamental microphysical entity in the language of microphysics. Q is an arbitrary phenomenal truth: perhaps the truth that someone is phenomenally conscious, or perhaps the truth that a certain individual (that is, an individual satisfying a certain description) instantiates a certain phenomenal property. $P\&\neg Q$ conjoins the former with the denial of the latter.

If Q is the truth that someone is phenomenally conscious, then $P\&\neg Q$ is the statement that everything is microphysically as in our world, but no-one is phenomenally conscious. In this version, $P\&\neg Q$ says that the world is a *zombie world*. If Q is the truth that a certain individual instantiates a certain phenomenal property, then $P\&\neg Q$ is the statement that everything is microphysically as in our world, but that it is not the case that the individual in question instantiates the relevant phenomenal property. In this case, it will suffice for the truth of $P\&\neg Q$ that the world is a zombie world, or simply that the individual in question is a zombie in a physically identical world. It will also suffice that the individual in question is an *invert*, who has an experience that differs slightly from the corresponding experience of the corresponding individual in our (physically identical) world.

The first premise of this argument asserts an epistemic thesis, about what can be conceived. The second premise steps from the epistemic thesis to a modal thesis, about what is possible. The third premise steps from the modal thesis to a metaphysical thesis, about the nature of our world.

The third premise is relatively uncontroversial. It is widely accepted that materialism has modal commitments. Some philosophers question whether materialism is equivalent to a modal thesis, but almost all accept that materialism at least *entails* a modal thesis. Here one can invoke Kripke's metaphor: if it is possible that there is a world physically identical to our world but phenomenally different, then after God fixed the physical facts about our world, he had to do more work to fix the phenomenal facts.

A familiar complication arises from the observation that physicalism about our world is compatible with the possibility of dualism in other worlds, and in particular is compatible with the possibility of a physically identical world that contains extra, nonphysical phenomenology. This

means that if Q is a negative truth about our world—say, the truth that no-one instantiates a certain phenomenal property—then materialism about our world is compatible with the possibility of $P\&\neg Q$. To finesse this point, we can stipulate that in the argument above, Q is a positive truth (one that holds in all worlds that contain a duplicate of our world; see Chalmers 1996, p. 40): if Q is a positive truth, then materialism is incompatible with the possibility of $P\&\neg Q$. Alternatively, we can conjoin P with a "that's-all" statement T, stating that the world is a *minimal* world that satisfies P (see Jackson 1998, p. 26). Then even when Q is a negative truth, materialism is not compatible with the possibility of $PT\&\neg Q$ (where PT is the conjunction of P and T).

The real work in the argument is done by the first and second premises. The second premise is particularly controversial, as there are a number of examples that have led many philosophers to deny that there is an entailment from conceivability to metaphysical possibility. To assess these premises, we need to understand the notion of conceivability.

2 Varieties of Conceivability

Conceivability is to be understood as an epistemic notion, defined in epistemological (and perhaps psychological) terms. To a first approximation, we can say that S is conceivable when S expresses a coherent hypothesis: one that cannot be ruled out a priori. To refine this understanding, it is useful to make some distinctions. (These distinctions are discussed at much greater length in Chalmers 2002.)

We can say that S is *prima facie* conceivable (for a subject) when that subject is unable to rule out the hypothesis expressed by S by a priori reasoning, on initial consideration. We can say that S is *ideally* conceivable when the hypothesis expressed by S cannot be ruled out a priori, even on ideal rational reflection. The main difference here is that prima facie conceivability is tied to a subject's contingent cognitive limitations, while ideal conceivability abstracts away from those limitations.

Some examples: (1) '2+2=5' is neither prima facie conceivable nor ideally conceivable; (2) Where S is a highly complex but provable mathematical truth, $\neg S$ will be prima facie conceivable for most subjects, but it is not ideally conceivable; (3) Where S is 'There is a flying pig', S is prima facie conceivable, and is almost certainly ideally conceivable.

The notions of conceivability discussed above are versions of *negative* conceivability which is defined in terms of what a subject can *rule out* through a priori reasoning. We can say that S is negatively conceivable when S cannot be ruled out through a priori reasoning. The two notions

above can then be seen as prima facie negative conceivability and ideal negative conceivability respectively.

It is also possible to define notions of *positive* conceivability, which is defined in terms of what subjects can form a positive conception of. We can say that S is positively conceivable when one can coherently imagine a situation in which S is the case. Where negative conceivability requires merely entertaining a hypothesis and being unable to rule it out, positive conceivability involves being able to form some sort of clear and distinct conception of a situation in which the hypothesis is true. One can then say that S is *prima facie positively conceivable* when a subject can imagine a situation that they take to be coherent and that they take to be one in which S is the case. And one can say that S is *ideally positively conceivable* when its prima facie positive conceivability cannot be defeated on ideal rational reflection (in particular, when arbitrary details can be filled in in the imagined situation without any contradiction revealing itself, and when ideal reflection reveals the imagined situation as one in which S is the case).

Traditional notions of conceivability (Descartes' clear and distinct conceivability, for example) are arguably varieties of positive conceivability rather than negative conceivability. At the same time, the notion of positive conceivability is more complex than that of negative conceivability, and a rigorous characterization of the notion requires saying much more about just what it is to imagine a situation, and so on. I characterize positive conceivability in more depth in Chalmers 2002. In this paper, the roles of positive and negative conceivability will often be interchangable (I will make clear when the difference is relevant), so the informal account above will suffice for present purposes. For much of the discussion one can focus on negative conceivability without much loss, but positive conceivability is available as an alternative if there turn out to be any problems with theses tied to negative conceivability.

Insofar as there is a gap between prima facie conceivability and ideal conceivability, it is ideal conceivability that is a better guide to possibility. This is especially clear in the case of prima facie negative conceivability: we have seen that the negation of a complex mathematical truth may be prima facie negatively conceivable, but it is not ideally conceivable and it is not possible. It is less easy to find cases of prima facie positive conceivability without ideal positive conceivability (see Chalmers 2002 for some potential cases), but insofar as there are such cases, there will be little reason to think that they are possible.

Correspondingly, some familiar purported counterexamples to the claim that conceivability entails possibility are really counterexamples to the claim that prima facie conceivability entails possibility. For example, it is sometimes said that both Goldbach's conjecture and its negation

are conceivable, while only one of them is possible. Here the relevant notion of conceivability is something like prima facie negative conceivability. There is no reason to believe that both Goldbach's conjecture and its negation are ideally conceivable, so there is no reason to think that this is a counterexample to the claim that ideal conceivability entails possibility. So from here onward, talk of "conceivability" simpliciter should always be understood to be talk of ideal conceivability (either positive or negative).

The other familiar class of purported counterexamples arises from Kripke's analysis of the necessary a posteriori. It is often said that sentences such as 'water is not H_2O' provide counterexamples to the claim that conceivability entails possibility: it is conceivable that water is not H_2O, but it is not metaphysically possible.

Here one has to be careful. There is a *sense* of 'conceivable' in which 'water is not H_2O' is not conceivable (given that water is H_2O in the actual world): in this sense, any conceivable situation in which it seems that water is not H_2O (a Twin Earth world, say) should better be described as a conceivable situation in which water is still H_2O, but in which there is watery stuff that is not H_2O. Using the term 'conceivable' this way, one might say that 'water is not H_2O' seems conceivable (or that it is prima facie conceivable, to one without relevant empirical knowledge), but that it is not really conceivable. We might call this sense of conceivability *secondary conceivability* (for reasons familiar from a two-dimensional analysis, and discussed in Chalmers 2002). Then the Kripkean cases are compatible with the claim that secondary conceivability entails metaphysical possibility. But at the same time, this claim is not very useful for present purposes, as whether a sentence is secondarily conceivable will typically depend on a variety of empirical factors, and an opponent might deny that zombies are secondarily conceivable, on the grounds that there is an a posteriori identity between phenomenal and physical properties. So a link between secondary conceivability and possibility does not offer an a priori route to conclusions about metaphysical possibility.

Instead, what is relevant here is *primary conceivability*: the sense in which 'water is not H_2O' can correctly be said to be conceivable. The notion of negative conceivability defined above is a sort of primary conceivability, as it is defined in terms of what can be ruled out a priori, and 'water is H_2O' cannot be established a priori. (One might define a distinct notion of negative secondary conceivability, but I will set that aside here.) One can likewise define a notion of positive primary conceivability, so that S is positively primarily conceivable when S can imagine a coherent situation that verifies S, where a situation verifies S when, under the hypothesis that the situation actually obtains, the subject should conclude that S. If the subject imagines a Twin

Earth situation with XYZ in the oceans and lakes, and assumes that the situation obtains in their own environment, then the subject should conclude that water is XYZ rather than H_2O. So 'water is not H_2O' is positively primarily conceivable, as well as negatively primarily conceivable.

Unlike secondary conceivability, matters of primary conceivability are plausibly in the a priori domain: whether S is primarily conceivable turns on matters of a priori reasoning. But primary conceivability does not entail metaphysical possibility: 'water is not H_2O' is primarily conceivable, but it is not metaphysically possible.

Still, there remains a link between primary conceivability and metaphysical possibility in these cases. When we conceive that water is not H_2O, we imagine (for example) a Twin Earth situation in which the watery liquid in the oceans and lakes is XYZ. This situation is metaphysically possible, so there is a sense in which our conceiving involves access to a possible world. Under the usual way of describing possible worlds, this world is not a world in which water is not H_2O. But the world still stands in a strong relation to the sentence 'water is not H_2O'. In particular, if we came to accept that our own world had the character of this world (with XYZ in the oceans and lakes), we should then endorse the claim 'water is not H_2O'.

This can be put in two-dimensional terms by saying that while the Twin Earth does not *satisfy* 'water is not H_2O' ('water is not H_2O' is not true of that world considered as counterfactual), the Twin Earth world *verifies* 'water is not H_2O' ('water is not H_2O' is true of that world considered as actual). Equivalently, we can say that while the *secondary intension* of 'water is not H_2O' is false at w, the sentence's *primary intension* is true there. To a first approximation, a world w verifies S (or S is true at w considered as actual, or the primary intension of S is true at w) when, if we came to accept that our own world is qualitatively like w, we should then endorse S. Strictly speaking, the worlds w that are relevant to primary intensions are *centered* worlds: worlds that come with a marked "center" consisting of an individual and time. When we consider a centered world w as actual, we consider the hypothesis that we are currently in the situation of the individual at the center. (For much more on these notions, see Chalmers 2004.)

We can say that when the primary intension of S is true at some centered world (i.e., when some centered world verifies S), S is *primarily possible*, or *1-possible*. When the secondary intension of S is true at some world (i.e., when some world satisfies S), S is *secondarily possible*, or 2-possible. Then 'water is not H_2O' is not 2-possible, but it is 1-possible.

The observation that sentences such as 'water is not H_2O' are conceivable but not possible, in these terms, comes to the claim that these sentences are primarily conceivable (1-conceivable) but are not secondarily possible (2-possible). So there is good reason to believe that 1-conceivability

does not entail 2-possibility. However, these cases are entirely compatible with a link between 2-conceivability and 2-possibility, and more importantly for present purposes, they are entirely compatible with a link between 1-conceivability and 1-possibility.

In fact, it is not hard to argue that all of the standard Kripkean a posteriori necessities ('heat is the motion of molecules', 'Hesperus is Phosphorus', and so on) have this structure. For each of these necessities, one might say that it's negation is conceivable but not possible, meaning that it is 1-conceivable but not 2-possible. But in each of these cases, the sentence in question is 1-possible. For example, 'heat is not the motion of molecules' is verified by a world in which something other than molecules causes sensations as of heat. 'Hesperus is not Phosporus' is verified by a world in which the objects visible in the morning and evening skies are entirely distinct. Furthermore, it is plausible that worlds such as these are just what one is conceiving of when one conceives that heat is not the motion of molecules, or that Hesperus is not Phosphorus. So in these cases, there remains a strong link between conceivability and metaphysical possibility.

To summarize, we have seen that the standard counterexamples to a conceivability-possibility link are accommodated by noting that (i) prima facie conceivability is an imperfect guide to possibility, and (ii) primary conceivability is an imperfect guide to secondary possibility. But (i) is entirely consistent with a link between ideal conceivability and possibility, and (ii) is entirely consistent with a link between primary conceivability and primary possibility. Putting the pieces together: all of these counterexamples are compatible with the thesis that ideal primary conceivability entails primary possibility.

There are two versions of this thesis, depending on whether one interprets the relevant sort of conceivability as positive or negative.

(CP+) Ideal primary positive conceivability entails primary possibility

(CP-) Ideal primary negative conceivability entails primary possibility.

CP- entails CP+, as ideal primary positive conceivability entails ideal primary negative conceivability. If S can be ruled out a priori, then no coherent imagined situation will verify S. It is not obvious whether or not CP+ entails CP-, as it is not obvious whether ideal primary negative conceivability entails ideal primary positive conceivability. That is, it is not obvious whether or not there is an S that cannot be ruled out a priori, but such that no coherent imagined situation verifies S. (In Chalmers 2002 I argue that there is no such S, so that ideal primary negative conceivability entails ideal primary positive conceivability.) So CP- is at least as strong as CP+ and is possibly somewhat stronger.

Most importantly for present purposes, however, both CP+ and CP- are compatible with all the familiar purported counterexamples to the conceivability-possibility link. Furthermore, it seems that there are no clear counterexamples to either thesis (though later in the paper, I will discuss some potential counterexamples that have been put forward). In particular, both theses are entirely compatible with the existence of Kripkean a posteriori necessities, so while existence of these necessities is often used to case doubt on conceivability-possibility theses, they cannot be used to cast doubt on CP+ or CP-.

So for now, I will take these theses as reasonable conceivability-possibility theses that might be used in mounting a refined conceivability argument against materialism. Later in the paper, I will return to the question of their truth.

3 A refined conceivability argument

Henceforth, unqualified uses of "conceivability" and "conceivable" should be understood as invoking ideal primary conceivability. I will often be inexplicit about whether positive or negative conceivability is involved. In effect, the argument forms below can be understood as generating two different arguments, depending on whether one understands conceivability as ideal primary positive conceivability or as ideal primary negative conceivability. For many purposes the distinction will not matter. When it does matter, I will be explicit.

Given the discussion above, one might try generating an anti-materialist argument by simply substituting primary possibility for metaphysical possibility in the original argument.

(1) $P\&\neg Q$ is conceivable

(2) If $P\&\neg Q$ is conceivable, $P\&\neg Q$ is 1-possible

(3) If $P\&\neg Q$ is 1-possible, materialism is false.

(4) Materialism is false.

On this reading, (1) and (2) are both plausible theses, but (3) is not obviously plausible. The reason is that materialism requires not the 1-impossibility of $P\&\neg Q$ but the 2-impossibility of $P\&\neg Q$. That is, materialism requires that it *could not have been the case* that P were true without Q being true. This is a subjunctive claim about ordinary metaphysical possibility, and so invokes 2-impossibility rather than 1-impossibility.

A materialist might reasonably question (3) by holding that even if there is a world w *verifying* $P\&\neg Q$, w might be a world with quite different ingredients from our own. For example, it might be that W does not instantiate true microphysical properties (those instantiated in our world), such as mass and charge, but instead instantiates quite different properties: say, pseudo-mass and pseudo-charge, which stand to mass and charge roughly as XYZ stands to $\mathrm{H_2O}$. Likewise, it might be that w does not lack true phenomenal properties, but instead lacks quite different properties: say, pseudophenomenal properties. If so, then the possibility of w has no bearing on whether true microphysical properties necessitate true phenomenal properties. And it is the latter that is relevant for materialism.

Still, it may be that the gap between 1-possibility and 2-possibility could be closed. In particular, when a statement S has the same primary intension and secondary intension, then a world will verify S iff it satisfies S, so S will be 1-possible iff it is 2-possible. If P and Q both have primary intensions that coincide with their secondary intensions, then so will $P\&\neg Q$, and we could run the following argument:

(1) $P\&\neg Q$ is conceivable

(2) If $P\&\neg Q$ is conceivable, $P\&\neg Q$ is 1-possible

(3) If $P\&\neg Q$ is 1-possible, $P\&\neg Q$ is 2-possible.

(4) If $P\&\neg Q$ is 2-possible, materialism is false.

(5) Materialism is false.

Here, the truth of (3) requires that both P and Q have primary and secondary intensions that coincide. In the case of Q, this claim is quite plausible. As Kripke noted, there does not seem to be the same strong dissociation between appearance and reality in the case of consciousness as in the cases of water and heat: while it is not the case that anything that looks like water is water, or that anything that feels like heat is heat, it is plausibly the case that anything that feels like consciousness is consciousness. So it is not clear that the notion of "pseudoconsciousness", something that satisfies the primary intension of 'consciousness' without being consciousness, is coherent. Likewise for other more specific phenomenal properties. So there is a strong case that the primary and secondary intensions of phenomenal terms coincide. (For more on this case, see Chalmers 2003b.)

However, in the case of P, this claim is less plausible. A materialist might reasonably hold that microphysical terms (such as 'mass' and 'charge') have primary intensions that differ from

their secondary intensions. In particular, it is plausible that the primary intensions of these terms are tied to a certain theoretical role. We might say that the primary intension of 'mass' picks out whatever property plays the mass-role (e.g. resisting acceleration in certain ways, being subject to mutual attraction in a certain way, and so on), and that the primary intension of 'charge' picks out whatever property plays the charge-role (e.g. obeying certain electromagnetic principles, being subject to attraction and repulsion in certain ways, and so on).

By contrast, one might reasonably hold that the secondary intension of microphysical terms is tied to the property that actually plays the role. For example, if property M plays the mass-role in the actual world, then one might hold that in any world in which mass is instantiated, mass is M. It follows that if there are worlds in which some *other* property M' plays the mass-role, then M' is not mass in that world (at best, it is pseudo-mass). If so, then the primary and secondary intensions of 'mass' will not coincide: the primary intension of 'mass' will pick out whatever plays the mass-role in such a world, but the secondary intension will not.

There are other views of the semantics and metaphysics of microphysical terms that may reject this argument for the distinctness of the primary and secondary intensions of 'mass'. In particular, the argument will not go through on views according to which it is necessary that mass is the property that plays the mass-role. (These include views on which 'mass' is a nonrigid designator whose secondary intension picks out different properties that play the role in some worlds, and views on which it is necessary that M is the property that plays the mass-role, where M is the property rigidly designated by 'mass'.) Still, the view sketched above is a quite reasonable view—more plausible than the alternatives, in my opinion—and it is the view that best grounds resistance to an inference from the 1-possibility of $P\&\neg Q$ to its 2-possibility. So we can suppose that the view is correct in order to see what follows.

On this view, a world may verify P without satisfying P. The secondary intension of P requires that certain specific properties such as mass, spin, and charge are distributed in a certain specific way across space-time, with appropriate causal and nomic relations among them. The primary intension of P requires only that whatever properties plays the mass-role, the spin-role, and charge-role are distributed in this way. If w is a world where these roles are played by properties other than mass, spin, and charge (we might say that they are played by "schmass", "schmin", and "schmarge"), which are otherwise distributed in the right way over space-time and have appropriate causal and nomic relations among them, then w will verify P, but it will not satisfy P. Here, we might say that the physics of W has the same *structural* profile as physics in the actual world, but that it has a different *intrinsic* profile, in that it differs in the intrinsic properties that fill this

structure. To verify P, a world must have the right structural profile, while to satisfy P, a world must have the right structural and intrinsic profile.

It follows that premise 3 is not guaranteed to be true. Because a world can verify P without satisfying P, it may be that $P\&\neg Q$ is 1-possible but not 2-possible. However, this requires that P and Q be related in a certain specific way. In particular, it requires that some worlds that verify P also verify $\neg Q$, while no worlds that satisfy P also satisfy $\neg Q$. This requires in turn that some worlds that have the same structural profile as the actual world verify $\neg Q$, while no worlds that have the same structural and intrinsic profiles as the actual world satisfy $\neg Q$. We can assume for the moment that the primary and secondary intensions of Q coincide. Then we can put all this by saying that the falsity of (3) requires that the structural profile of physics in the actual world does not necessitate Q, but that the combined structural and intrinsic profiles of physics the actual world do necessitate Q.

This idea—that the structural properties of physics in the actual world do not necessitate the existence and/or nature of consciousness, but that the intrinsic properties of physics combined with the structural properties do—corresponds to a familiar view in the metaphysics of consciousness. This is the view that I have elsewhere called *Russellian monism* (or type-F monism, or panprotopsychism). On this view, consciousness is closely tied to the intrinsic properties that serve as the categorical bases of microphysical dispositions. Russell and others held that the nature of these properties is not revealed to us by perception (which reveals only their effects) or by science (which reveals only their relations). But it is coherent to suppose that these properties have a special nature that is tied to consciousness. They might themselves be phenomenal properties, or they might be *protophenomenal* properties: properties that collectively constitute phenomenal properties when organized in the appropriate way.

Russellian monism is an important view on the mind–body problem. I think that it is certainly not ruled out by the conceivability argument and by related arguments. If Russellian monism is true, then when we conceive of zombies, we hold fixed the structural properties of physical systems in the actual world, but not their intrinsic properties (which are protophenomenal properties). If we consider these intrinsic protophenomenal properties to be physical properties, then Russellian monism will qualify as a form of physicalism. But because it relies on speculation about the special nature of the fundamental properties in microphysics, it is a highly distinctive form of physicalism that has much in common with property dualism, and that many physicalists will want to reject.[1]

[1] A related position arises from views on which laws of nature are necessary (e.g. Shoemaker 19xx), and on which there is a lawful connection between physical properties and phenomenal properties in our world. Such a view may

In any case, we can now close the loophole in the previous argument as follows:

(1) $P\&\neg Q$ is conceivable

(2) If $P\&\neg Q$ is conceivable, then $P\&\neg Q$ is 1-possible

(3) If $P\&\neg Q$ is 1-possible, then $P\&\neg Q$ is 2-possible or Russellian monism is true.

(4) If $P\&\neg Q$ is 2-possible, materialism is false.

(5) Materialism is false or Russellian monism is true.

This argument is valid. I discussed the case for premises (1), (2), and (4) earlier, and I have just now argued for premise (3). I think that (5) is the proper conclusion of the conceivability argument. For the reasons given above, such arguments (and also related arguments such as the knowledge argument and the property dualism argument) cannot exclude Russellian monism, and Russellian monism is arguably a form of physicalism, if a distinctive and radical kind. So the possibility of Russellian monism needs to be explicitly acknowledged as an option in the conclusion.

A couple of minor notes on the argument. First, to be fully explicit, the argument might take the truth of Q as a premise. If Q were false, the ground for accepting (4) would collapse. In the less explicit version of the argument above, we can consider the truth of Q part of the case for accepting premise (4). In fact, for reasons given earlier, the case for (4) requires that Q is a *positive* truth about consciousness. Alternatively, one can remain silent on whether Q is a positive or negative truth, and handle this matter by conjoining P with a "that's-all" clause asserting that the world is a minimal world in which P (or equivalently, by building such a that's-all clause into P).

Second, it is worth noting that (contrary to a common supposition), the assumption that Q has the same primary and secondary intensions is not necessary for the argument for (5) to go through. To see this, we can consider the version of the argument where we adjoin a "that's-all" clause to P. From (1) and (2), we can derive the conclusion that there is a minimal world verifying P in

hold that it is essential to physical properties that they have this nomic profile, so that there is no world satisfying $P\&\neg Q$. Some versions of this view will deny (2), and are discussed later in the paper. But other versions may accept (2), holding that there is a world verifying $P\&\neg Q$, but will hold that it involves distinct "schmysical" properties that lack this nomic profile. The resulting view resembles Russellian monism in some respects, but differs from the usual form in taking the connection between physical and phenomenal properties to be nomic in the first instance. Because it turns on this nomic connection, this view does not provide any loophole for materialism: at best, it yields a version of property dualism on which the laws of nature connecting physical and phenomenal properties are necessary. See objection 12 in section XX for more on related issues.

which the primary intension of Q is false. If P has the same primary and secondary intensions, then this world will be a minimal P-world in which the primary intension of Q is false. This world must differ from our world, because the primary intension of Q is true in our world. (There is a small loophole here arising from the possibility that this world differs merely in the location of the center of the relevant centered world. I discuss this loophole in section 5.) It follows that there is a minimal P-world that is not a duplicate of our world, so that physicalism is false of our world. It could be that strictly speaking physicalism will be true of *consciousness*, because P necessitates Q, but physicalism will be false of properties closely associated with consciousness, namely those associated with the primary intension of Q. We might think of this sort of view as one on which phenomenal properties are physical properties that have non-physical properties as modes of presentation.

Alternatively, if P has different primary and secondary intensions, then by the reasoning given in the earlier discussion of premise (3), one can conclude that either there is a minimal world satisfying P in which the primary intension of Q is false (which again entails the falsity of physicalism), or that the primary intension of Q is necessitated by the structural and intrinsic profiles of physics in our world, but not by the intrinsic profiles alone. This view can be considered another form of Russellian monism, in that the intrinsic properties of physics in our world are crucial for constituting the properties associated with the modes of presentation of consciousness. So if Q has a distinct primary and secondary intensions, then one will have to formalize premises (3) and (4) somewhat differently, but the argument for (5) will still work just as well.

This completes the exposition of the two-dimensional argument against materialism. In what follows I will address some objections to the argument that arise from the recent literature. I have already said what needs to be said about premises (3) and (4), and I do not anticipate significant objections to them (with one exception that I will discuss). I will discuss some objections to premise (1) and some hard-to-classify objections relatively briefly, and then I will focus on objections to the crucial premise (2).

4 Objections to Premise (1)

Premise (1) says that $P\&\neg Q$ is conceivable: that is, that we can conceive a world physically identical to ours containing zombies and/or inverts. It is worth noting that such a world need not be an entire zombie world. One might worry that there are complications associated with imagining *oneself* to be zombie, at the center of a centered world. For the purpose of the argument, it usually

suffices to conceive of a physically identical world in which some other being, corresponding to a conscious being in this world, is a zombie or an invert.

The premise can be understood as invoking ideal primary negative conceivability or ideal primary positive conceivability. The first version makes a somewhat weaker claim (positive conceivability entails negative conceivability, but the reverse is not obviously the case), and is slightly more straightforward. In this version, the premise simply says that $P\&\neg Q$ cannot be ruled out a priori: that a priori reasoning cannot rule out the hypothesis that P obtains and that someone else is a zombie, for example. The second version makes the somewhat stronger claim that we can (in principle) clearly and distinctly imagine a situation in which P holds and in which someone is a zombie or an invert, where the hypothesis that this situation obtains cannot be ruled out by a priori reasoning. Of course, any difference in strength here will be balanced by an inverse difference in strength for the corresponding versions of premise (2).

I discussed the strong prima facie case for these claims in Chalmers (1996), and will not try to recapitulate it here. The zombie hypothesis is at least prima facie coherent and imaginable. To reject the premise, one needs to find something that undermines the prima facie coherence and imaginability, such as some sort of a priori incoherence, contradiction, or unimaginability in the hypothesis that emerges on reflection. A detailed defense of the premise involves arguing that no candidate for this sort of undermining can succeed. Here I will consider various suggestions.

(1) *Analytic functionalism*. One way to reject the conceivability premise would be to accept an analytic functionalist view of consciousness, on which what it means to say that someone is conscious is that they have a state that functions in an appropriate way in their cognitive system and their behavior. If consciousness were defined in this way, then any functionally identical being would be conscious by definition, so zombies would be ruled out. Analytic functionalism about consciousness is widely rejected, however, and I have argued against it at length elsewhere (see e.g. the discussion of type-A materialism in Chalmers 2003a), so I will not argue against it further here.

(2) *Prima facie but not ideal conceivability*. It is natural to suggest that zombies may be prima facie but not ideally conceivable (see van Gulick 2000 and Worley 2003 for suggestions along these lines). This corresponds to the position that I call type-C materialism in Chalmers (2003a), according to which there is a prima facie epistemic gap between physical and phenomenal truths that may be closed on ideal reflection, perhaps because of new discoveries about physical processes, or perhaps because of novel reasoning. To very briefly summarize the argument there: to render zombies incoherent even on ideal reflection requires some sort of conceptual link between

physical and phenomenal concepts. Given that physics and physical concepts are all structural-dynamical in character (and new scientific developments are unlikely to change this, although see option (3)), phenomenal concepts must have a character that is linked to structural-dynamic concepts in an appropriate way. Upon examination, the only candidate that is remotely tenable is the hypothesis that phenomenal concepts are functional concepts. But we have already seen that there is good reason to reject that view.

A related objection (e.g. Bailey forthcoming) holds that arguments from ideal conceivability are toothless, as non-ideal creatures such as ourselves cannot know whether or not a given statement is ideally conceivable. I think that there is no reason to accept this claim. Although we are non-ideal, we can know that it is not ideally conceivable that 0=1, and that it is ideally conceivable that someone exists. We know that certain things about the world (say, all philosophers are philosophers) are knowable a priori, and that certain things about the world (say, that there is a table in this room) are not so knowable, even by an ideal reasoner. Likewise, reasoning of the sort above gives us very good reason to think that there is no a priori entailment from physical to phenomenal truths and that zombie hypotheses are conceivable, even for an ideal reasoner.

(3) *Expanding the conception of the physical*. Stoljar (2001) argues that whether zombies are conceivable depend on how we understand physical properties. When we conceive of zombies, we hold fixed *t-physical* properties, or properties as characterized in physical theory. But we may not hold fixed *o-physical* properties, or the underlying properties of physical objects, including the intrinsic properties that play the roles characterized in physical theory. So the position is left open that o-physical zombie duplicates are both inconceivable and impossible: if we adequately conceived of the o-physical properties, we would see that any o-physical duplicate of a conscious being is conscious. And if we expanded P to include hypothetical o-physical concepts of these intrinsic properties (characterizing their intrinsic nature), then $P\&\neg Q$ would be inconceivable. I think that this is an important position, but it is clearly a version of Russellian monism, so I take it to be compatible with the overall argument given above.

(4) *Zombies presuppose epiphenomenalism*. John Perry (2001) has suggested that the assertion that zombies are conceivable presupposes epiphenomenalism, the thesis that consciousness plays no causal role. If we thought that consciousness plays a causal role, we would not find zombies conceivable. In Chalmers (2004b) I argue that this claim is simply false. Its falsity is demonstrated by the fact that many non-epiphenomenalist views are compatible with the conceivability of zombies. For example, many type-B materialists (such as Hill, Loar, Papineau, and so on) accept that zombies are conceivable, and that consciousness plays a causal role. These philoso-

phers deny that zombies are possible, but that is a different stage of the argument. Furthermore, the Russellian monist view is a non-epiphenomenalist view that we have seen is compatible with the conceivability of zombies in the relevant sense. Finally, even Cartesian interactionist dualism, on which consciousness certainly plays a causal role, is compatible with the conceivability (and possibility) of zombies. On such a view, physically identical beings without consciousness will presumably have large causal gaps in their functioning (or else will have some new element to fill those gaps), but there is nothing obviously inconceivable about such causal gaps. (For more on this, see Chalmers 2004b.)

(5) *The judgments of zombies*. It is an admittedly strange feature of zombies that they appear to make *claims* about consciousness that are indistinguishable from the claims of conscious beings, and that they even make parallel *judgments*, if judgments are functionally understood. I called this "the paradox of phenomenal judgment", and argued that while it is strange (many strange things happen in other possible worlds), it does nothing to undermine the coherence of zombies. Thomas (1998) argues that we have to choose between saying that zombie's claims are true, meaningless, insincere, or mistaken, and that each of these options is untenable. Others (Kirk 1999, Lynch forthcoming) have argued that the coherence of zombies undermines our *knowledge* of consciousness, since we will then be unable to distinguish ourselves from zombies. In Chalmers (2003b) I argue that an appropriate account of the content and epistemology of our phenomenal judgments shows that there is no problem here.

(6) *Zombies are not positively conceivable*. Some (e.g. Ashwell 2003, Marcus 2004) have argued that while zombies may be negatively conceivable, they are not positively conceivable. To positively conceive another creature, one must conceive of some combination of physical processes and conscious states. But conceiving of physical processes clearly does not suffice to conceive of a zombie (since conscious beings may have the same processes), and in conceiving of a zombie it is out of the question to conceive of conscious states. So we can form no positive conception of a zombie.

In response, I think it is clear that when we conceive of zombies we conceive of the *absence* of consciousness. There is no more problem with clearly and distinctly imagining a situation in which there is no consciousness than in imagining a world in which there are no angels, or in imagining a world with one particle and nothing else. The argument here appears to require the idea that absences are never positively conceivable, or at least that to positively conceive an absence always requires conceiving something else in its place. But the cases above suggest that such a claim is clearly false. The claim may turn on misinterpreting the notion of positive conceivability:

presumably there is a sense in which conceiving of an absence is negative conceiving, but this is not the sense that is relevant here.

A general point about the arguments above: Many of these arguments, especially 4-6, turn on considerations that are specific to the conceivability of zombies, and do not apply to the conceivability of inverts and the like. One should keep in mind that for the anti-materialist argument, one does not need to consider beings as remote from us as zombies, or even as remote as full-scale inverts. It suffices if we can conceive of a being whose conscious experience is for just a moment slightly different from that of an actual physical duplicate's: perhaps they experience a slightly different shade at a point in the background of their visual field. Any problems that are specific to zombies then will not apply.

5 Hard-to-Classify Objections

Here I consider three objections to the argument that are not straightforwardly classifiable as objections to premise (1) or premise (2).

(1) *The conditional-concepts objection.*

David Braddon-Mitchell, John Hawthorne, and Robert Stalnaker have independently put forward versions of the following intriguing objection. The conceivability of zombies is compatible with materialism, if the concept of consciousness has a certain conditional structure. Suppose that 'consciousness' functions as follows: if dualism is true in the actual world (that is, if a relevant sort of nonphysical property is instantiated), then 'consciousness' picks out a nonphysical property; and if physicalism is true in the actual world (that is, if no relevant nonphysical property is instantiated but a relevant sort of physical property is instantiated), then 'consciousness' picks out a physical property. Then if dualism is actually true, we should accept that zombies are possible. And it is at least conceivable that dualism is actually true, in that we cannot rule out the truth of dualism a priori. It follows that it is conceivable that zombies are possible. But this conceivability is entirely compatible with the claim that dualism is actually false, so that in fact zombies are impossible.

In response: Let us say that S is *meta-conceivable* when it is conceivable that S is possible. Then this objection shows that the *meta-conceivability* of zombies may be compatible with materialism. But the relevant premise of the conceivability argument was not that zombies are meta-conceivable, but that zombies are conceivable. That is, the premise (at least in the negative

conceivability version) asserts that we cannot rule out a priori that $P\&\neg Q$ obtains in the actual world, not that we cannot rule out a priori that $P\&\neg Q$ obtains in some possible world. But there are good reasons to accept this premise, and the considerations above does nothing to undermine them. So these considerations do not threaten the argument.

Hawthorne and Stalnaker do not discuss the claim that zombies are conceivable (as I am understanding that claim here), and they say nothing to undermine either that claim or the claim that conceivability in this sense entails possibility, so their discussion leaves all the premises in the argument above intact. Braddon-Mitchell, by contrast, makes clear that he accepts premise 2 but denies premise 1. He says that the apparent conceivability of zombies is just a "shadow" of their metaconceivability: that is, we are apt to conflate the claim that zombies are conceivable with the claim that they are metaconceivable, and we mistakenly take the good reasons for the latter to be reasons for the former. He does not say anything to argue directly that zombies are inconceivable, however. On the face of it, even once we distinguish the two notions, the grounds for saying that zombies are conceivable (e.g., that there is no a priori contradiction in the idea) are as strong as ever.

These proponents could suggest that the conditional account predicts that zombies are inconceivable, and that it has independent support. At least on Braddon-Mitchell's version of the account (Hawthorne and Stalnaker are not explicit about this), it is an a priori conceptual truth that if PT is the case, then Q is the case, as he takes the conditional account to be an a priori conceptual analysis of phenomenal concepts. Braddon-Mitchell supports this claim by arguing that if an oracle told us that the world is purely physical, we would respond by accepting physicalism about phenomenal consciousness, rather than denying that it exists. (Hawthorne makes a similar observation.)

In response: I think that although the observation about our reaction to oracles is correct, it gives no support to the conditional concepts analysis. As Alter (forthcoming) has noted, the observation is grounded in the fact that we are more certain that we are conscious than we are of any philosophical thesis about consciousness. And our certainty that we are conscious is plausibly a posteriori knowledge, justified by experience. If so, then (as Alter argues), our inference from PT to Q in the oracle situation is partly grounded in this a posteriori knowledge, rather than being a priori. So the oracle observation does not support the thesis that it is a priori that if PT is the case, Q is the case. Furthermore, that thesis is undermined by both the zombie intuition itself and the knowledge-argument intuition that someone who knows PT would not thereby be in a position to know (through a priori reflection) what it is like to see red. So we have good reason to

reject the conditional analysis of phenomenal concepts.

(2) *The zombie parity objection*.

Balog (1999) has suggested that there must be something wrong with the argument, on the grounds that if zombies are possible, they could make a parallel argument to the conclusion that materialism is false in their world. This conclusion would be incorrect, as materialism is true in their world. So something must be wrong with the zombie's argument. As the argument is valid, it must have a false premise. But Balog argues that if the premises of the original argument are true, then the premises of the zombie's argument are true. It follows that one of the premises of the original argument is false. Balog locates the problem in the equivalent of premise 2, the inference from conceivability to possibility, but this specific claim is not needed to undermine the original argument.

In response, I think it is not correct that if the premises of the original argument are true, the premises of the zombie's arguments are true. A tacit premise of the original argument is Q itself, stating for example that someone is phenomenally conscious. (As discussed earlier, this could either be built in as an explicit premise, or it can function as part of the support for premise 4.) The corresponding premise of the zombie's argument will be its claim 'Someone is phenomenally conscious'. I think that this claim is false. For example, in a debate between zombie realists and zombie eliminativists, the zombie eliminativists (who say 'No-one is phenomenally conscious') are correct. (For support for this claim, based in an analysis of phenomenal concepts, see Chalmers 2003b.) If this is right, then the zombie's argument has a false premise where the original argument has a true premise, and the failure of the zombie argument does nothing to undermine the original argument.

(3) *The indexical objection*.

A fairly common move in response to the conceivability argument (see e.g. Ismael 1999, Perry 2001) is to suggest that it is undermined by an analogy with indexicals. Even if I have complete objective knowledge of the world, I may lack indexical knowledge, such as the knowledge that it is 8pm now. So it is conceivable that all those objective truths obtain and that it is not 8pm now. But it does not follow that the fact that it is 8pm now is some ontologically further fact about the world, so that materialism is false. Likewise, it does not follow from the conceivability of zombies that materialism is false.

One might take this to be an objection to premise (2), the inference from conceivability to possibility. But in fact the point is compatible with premise (2). Premise (2) says that if $P\&\neg Q$ is conceivable, it is 1-possible. The analog version for time says that if $P\&\neg N$ is conceivable

(where N is "It is 8pm now"), then $P\&\neg N$ is 1-possible. But $P\&\neg N$ *is* 1-possible, in that there is a centered world verifying $P\&\neg N$: a world physically identical to our world, with the center placed somewhere else. In terms of the current argument, the objection is rather that from the fact that there is a *centered* world verifying $P\&\neg N$, it does not follow that there is an *uncentered* world satisfying $P\&\neg N$ (or satisfying anything in the vicinity), and such an uncentered world is what is needed for materialism to be false. By analogy, something similar goes for $P\&\neg Q$. So if we accept the analogy, we should deny premise (3).

In my main argument for premise (3) above, I assumed the thesis that Q has the same primary and secondary intension. If that thesis is correct, the indexical objection will not apply, as centering is irrelevant to the intensions of such sentences. If there is a centered world verifying $P\&\neg Q$, then the corresponding uncentered world will also satisfy $\neg Q$ (and any difference between the primary and secondary intensions of P will lead at worst to Russellian monism). However, a proponent of the indexical objection may hold that phenomenal concepts are themselves indexical concepts, which have distinct primary and secondary intensions (for example, the primary intension of "I" picks out the individual at the center of a centered world, while the secondary intension of "I" picks out the same individual in all worlds). If so, then Q will have distinct primary secondary intensions. When I considered this possibility earlier, I noted that the argument still goes through but that there is a small loophole in the argument due to centering. In effect, the proponent of the indexical objection is exploiting this loophole in order to resist the argument.

To make considerations about indexical concepts maximally explicit, one can modify the formal argument. The modified argument invokes the sentence PTI, a conjunction of P, T, and I. Here P and T are as before (the "that's-all" truth T is built in explicitly here for completeness). I is a conjunction of relevant indexical truths, such 'I am x', 'Now is y', and so on, where x and y are descriptions picking out unique individuals and times (see Chalmers and Jackson 2001 for more on this). The modified argument runs as follows:

(1) $PTI\&\neg Q$ is conceivable

(2) If $PTI\&\neg Q$ is conceivable, then $PTI\&\neg Q$ is 1-possible

(3) If $PTI\&\neg Q$ is 1-possible, then $PT\&\neg Q$ is 2-possible or Russellian monism is true.

(4) If $PT\&\neg Q$ is 2-possible, materialism is false.

(5) Materialism is false or Russellian monism is true.

Here premise (2) is an instance of the general principle discussed earlier, and premise (4) is as before. The case for the modified premise (3) is more or less the same as that for the earlier premise (3). The main difference is this version removes the residual loophole in the case for premise (3) at the end of section 3, after we relaxed the assumption that Q has the same primary and secondary intensions. That loophole involved the suggestion that a centered world $w*$ verifying $PT\&\neg Q$ need not differ in any objective respect from the actual centered world W, since it might differ only in the location of the center. That loophole is now removed: if both w and $w*$ verify PTI, then they cannot differ only in the location of the center, as I specifies that location.[2] So w and $w*$ will differ in objective respects despite both verifying PT, and the rest of the argument will go through as before.

As for the modified premise (1), this premise follows from the observation that adding indexical information does not close the epistemic gap between PT and Q. For example, I can coherently conceive that PT is the case, that I am x, that now is y, and that someone else is a zombie. (Note that this is a disanalogy with indexical cases: $PTI\&\neg N$ is not conceivable, where N is a claim about what time it is now.) Likewise, if we give Mary in her black-and-white room the same sort of indexical information, she will still not be in a position to know what it is like to be red. The epistemic gaps are just as wide as they were before. So the modified premise (1) is as plausible as the earlier version.

At this point, an opponent may suggest that phenomenal concepts are primitive indexical concepts, quite independent of other indexical concepts such as *I* and *now*. For example, phenomenal concepts might involve a primitive phenomenal indexical 'E', which picks out a phenomenal state of the experiencer in a way analogous to that in which 'I' and 'now' pick out the speaker and time. If so, then the full indexical truth I must include claims of the sort 'E is such-and-such', where the right-hand-side is an independent specification of the referent. And the opponent might suggest that this right-hand-side could specify a physical or functional property. That is, the indexical truth I might include a claim of the form 'E is PP', where PP picks out a physical or functional property instantiated by the subject. (Hawthorne 2002 makes a suggestion along these lines.) But if so, $PTI\&\neg Q$ may well be contradictory. Say that Q is 'Someone instantiates E'. P will entail that 'Someone instantiates PP', and in conjunction with I will entail 'Someone instantiates E'. If so, then $PTI\&\neg Q$ is not conceivable after all.

[2] Strictly speaking, there is an exception for cases of wholly symmetrical worlds where individuals cannot be picked out by any unique objective descriptions. But there is no reason to think that the actual world is such a world, and any such symmetries play no role in the judgment that $PTI\&\neg Q$ is conceivable.

I think that this interesting suggestion founders at the first stage: there is good reason to believe that our core phenomenal concepts are not indexicals. I think it is plausible that there are indexical phenomenal concepts such as E, which we might gloss as *this phenomenal property*, picking out a phenomenal property that the subject is currently having. But as I argue in Chalmers (2003b), these concepts are distinct from our core phenomenal concepts, such as the pure phenomenal concept of phenomenal redness. Crucially, any identity claim involving pairs of such concepts is cognitively significant. For example, the thought *this phenomenal property is R*, involving an indexical concept and a phenomenal concept of phenomenal redness, is as cognitively significant as a thought such as *this shape is circle*, involving an indexical and a geometric concept of circularity. If so, the phenomenal concept is distinct from the indexical concept.

One can also make a direct case against any analysis of phenomenal knowledge as indexical or demonstrative knowledge, as follows. The epistemic gaps associated with indexicals always disappear from an objective perspective. Say that I am physically omniscient, but do not know whether I am in the USA or Australia (we can imagine that there are appropriate qualitative twins in both places). Then I have am ignorant of the truth of 'I am in Australia', and discovering that I am in Australia will constitute new knowledge. But if someone else is watching from the third-person point of view and is also physically omniscient, they will have no corresponding ignorance concerning whether I am in Australia: they will know that A is in Australia and that B is in the US, and that is that. There is no potential knowledge that they lack: from their perspective, they know everything there is to know about my situation. So my ignorance is essentially indexical, and evaporates from the objective viewpoint. The same goes for indexical ignorance concerning what time it is: when I am ignorance of the truth of 'It is 3am now', a physically omniscient historian in later years will have no corresponding ignorance about whether it was 3am then. Demonstrative ignorance concerning what "this" is works in a similar way. In all these cases, the ignorance disappears from the objective viewpoint. An objectively omniscient observer can know everything there is for them to know about my situation, and there will be no doubts left for them to settle.

The epistemic gaps associated with phenomenal knowledge behave quite differently. Consider Mary in her black-and-white room, ignorant of what it will be like for her to see red for the first time. In this case, a physically omniscient observer may have precisely analogous ignorance: even given his complete physical knowledge, he may have no idea what it will be like for Mary to see red for the first time. So this ignorance does not evaporate from the objective viewpoint. The same goes even more strongly for knowledge of what it is like for others to see red. For any observer,

regardless of their viewpoint, there will be an epistemic gap between complete physical knowledge and this sort of phenomenal knowledge.

This suggests strongly that phenomenal knowledge is not a variety of indexical or demonstrative knowledge at all. Rather, it is a sort of objective knowledge of the world, not essentially tied to any viewpoint. If so, then any analysis of phenomenal concepts as indexical concepts will fail.

(4) *Objections to two-dimensional semantics*

One general sort of resistance to the two-dimensional argument comes from resistance to the two-dimensional framework as I have set it out. If one does not think that the notion of a primary intension is coherent, then most of the premises of the argument will not even be meaningful, let alone true. This sort of objection is too broad in scope to be covered in this paper; see Chalmers (2006b) for some replies to objections. Here, I will just note that much prominent resistance to two-dimensional semantics is really resistance to the claims that the intensions it postulates are the semantic content of utterances, or that they are what is said, or that they are what is ascribed in propositional attitude ascriptions, or some such similar thesis. But the arguments in this paper need no such thesis. All they need is that sentences or utterances can be associated with primary intensions in the way I describe. Nothing turns on whether this association yields a sort of "semantic content", can be used for attitude ascriptions, and so on. All that matters is that the relation of verification between worlds and utterances is well-defined (and then, of course, that it makes the premises true).

For example, a critical paper by Bealer (2002) focuses all of its arguments on the semantic theses, and gives no argument at all against the claim that statements can be associated with primary intensions or against the claim that they satisfy the CP theses. Something similar applies to a book on two-dimensionalism by Soames (2004), which is mostly devoted to arguing against certain two-dimensional analyses of attitude ascriptions, and spends very little time arguing against the underlying association of statements with primary intensions. Setting arguments of this sort aside, the majority of other arguments against two-dimensionalism are really arguments against specific theses such as the associated conceivability-possibility premise, and are addressed below. So henceforth I will assume the coherence of the framework, and will defend the specific theses.

6 Objections to Premise (2)

Finally, we come to the crucial premise (2). It is this premise that bridges the epistemic and modal domains, and it is this premise and associated principles that have attracted the most in-depth

philosophical discussion.

Premise (2) says that if $P\&\neg Q$ is conceivable, $P\&\neg Q$ is 1-possible. This premise can be seen as an instance of the general conceivability-possibility thesis CP:

CP: If S is conceivable, S is 1-possible.

Here, "conceivability" should be understood as ideal primary conceivability, of either the negative or positive variety (I always take "ideal primary" as understood from here on). The two versions of the thesis that result are equivalent to theses CP- and CP+, discussed earlier. Thesis CP- is equivalent to the claim that if $\neg S$ is not a priori, S is 1-possible. The positive version CP+, holding that if S is positively conceivable, S is 1-possible, is somewhat weaker than the negative version, as positive conceivability entails negative conceivability but the reverse is not obviously the case. Much of my discussion will apply equally to both CP+ or CP-, so I will often just speak of CP, except where the distinction is relevant.

The case for premise (2) largely derives from the case for CP, and from here on I will mostly focus on the general principle rather than the specific premise. Of course if it turns out that the general principle needs to be restricted to a certain class of sentences to be plausible, then the question will arise as to whether $P\&\neg Q$ falls into that class.

Why believe CP? In the first instance, the thesis is plausible because there are no clear counterexamples to it. Principles linking conceivability and possibility have been widely accepted in the history of philosophy, but have more recently been questioned because of various counterexamples, such as the Goldbach case (both Goldbach's conjecture and its negation are conceivable but only one is possible) and especially the Kripke cases ('Hesperus is not Phosphorus' is conceivable but not possible). But CP accommodates these examples straightforwardly, with the idealization accommodating Goldbach cases, and the primary/secondary distinction accommodating Kripke cases. If it handles these cases, then the central sources of resistance to conceivability-possibility principles is undermined. But of course, there may be other possible sources of resistance.

7 Strong Necessities

Before proceeding, it is useful to clarify CP by making clear what a counterexample to it would involve.

According to the two-dimensional analysis, ordinary Kripkean a posteriori necessities such as 'water is $\mathrm{H_2O}$' and 'Hesperus is Phosphorus' have a necessary secondary intension but a con-

tingent primary intension. That is, such statements are 2-necessary but 1-contingent: there are centered possible worlds (a Twin Earth world, or a world with distinct morning and evening stars) that verify their negations. When S is an a posteriori necessity of this sort, with a contingent primary intension, we might say that S is a *weak a posteriori necessity*.

By contast, we can say that an a posteriori necessity is a *strong a posteriori necessity*, or just a *strong necessity*, iff S has a necessary primary intension. Strong necessities are a posteriori necessities that are verified by all centered metaphysically possible worlds. It is not easy to give examples of strong necessities, as all of Kripke's a posteriori necessities appear to be weak necessities. But I will discuss some putative candidates in what follows.

It is easy to see that CP- is equivalent to the thesis that there are no strong necessities. If S is negatively conceivable but not 1-possible, then $\neg S$ will be a strong necessity. If S is a strong necessity, then $\neg S$ will be negatively conceivable but not 1-possible.

Insofar as CP+ is potentially weaker than CP-, the relationship between CP+ and the thesis that there are no strong necessities is not as clear. Certainly any counterexample to CP+ will yield a strong necessity, but the reverse is not obviously the case. To handle this, we might define two classes of strong necessities, according to whether they provide counterexamples to CP+ or merely to CP-. Let us say that a *negative strong necessity* is a statement S such that S is 1-necessary and 2-necessary but $\neg S$ is negatively conceivable. The latter condition is equivalent to the requirement that S is not a priori, so negative strong necessities are equivalent to strong necessities as defined above. A *positive strong necessity* is a statement S such that S is 1-necessary and 2-necessary while $\neg S$ is positively conceivable. Then all positive strong necessities are negative strong necessities, but the reverse is not trivially the case. CP- and CP+ are then equivalent to the theses that there are no negative strong necessities and that there are no positive strong necessities respectively.

What would a strong necessity involve? To get an idea, consider a philosophical view on which it is metaphysically necessary that an omniscient being (e.g. God) exists, but on which it is not a priori that such a being exists. Then according to this view, 'An omniscient being exists' (or O) is an a posteriori necessity. Like all a posteriori necessities, O is 2-necessary, and $\neg O$ is negatively conceivable (and also positively conceivable, if we add the plausible claim that it is positively conceivable that there is no omniscient being). If O were an ordinary a posteriori necessity, then O would be 1-contingent: there would be a metaphysically possible world verifying $\neg O$. But if there is no omniscient being, then it seems that there is no such world. 'There is an omniscient being' does not seem to have any difference in its primary and secondary intensions, so if a world

satisfies O, it verifies O. So given that O is 2-necessary, O is 1-necessary. It follows that if this philosophical view is correct, then O is a strong necessity: it is at least a negative strong necessity, and given the positive conceivability claim above, it is a positive strong necessity.

One could put the matter by saying that there is an epistemically possible *scenario* verifying $\neg O$, but no metaphysically possible *world* verifying $\neg O$. Here a scenario can be understood as corresponding to a maximal a priori coherent hypothesis, in the way that worlds correspond to maximal metaphysically possible hypotheses. (I give a formal treatment of scenarios in Chalmers (2004a, forthcoming), but here I will leave the notion intuitive. (One might call this sort of scenario a *negative scenario*, since it corresponds to a maximal negatively conceivable hypothesis. One could also define a *positive scenario* so that it corresponds to a maximal positively conceivable hypothesis.) The notion of scenarios is not defined in terms of metaphysical possibility, and in particular it is not assumed that scenarios correspond to metaphysically possible worlds. But nevertheless it is plausible that there is an intimate relationship.

For any a posteriori necessity S, there will be a scenario verifying $\neg S$. For example, as 'water is H_2O' is not a priori, there will be a scenario verifying 'water is not H_2O'. That is, there will be some maximal a priori coherent hypothesis H (perhaps involving the assumption that the watery stuff is made of XYZ, and so on) such that if we accept H, we should accept 'water is not H_2O'. For ordinary a posteriori necessities, these scenarios will correspond closely to centered metaphysically possible worlds, so that there will be a centered world verifying $\neg S$. For example, 'water is not H_2O' is verified by a centered XYZ-world, where the individual at the center is and has always been surrounded by clear, drinkable XYZ in the oceans and lakes. There is little reason to doubt that such a world is metaphysically possible, and there is an intuitive sense in which it qualitatively matches the scenario that we imagine when we entertain the hypothesis that water is not XYZ.[3]

(I will leave the notion of correspondence (or of qualitative matching) between scenarios and worlds at an intuitive level here, as it is not needed to play a formal role in the arguments. One might formalize it by requiring that scenarios be described in a limited vocabulary consisting of semantically neutral terms and indexicals, in the form D&I, where D is an semantically neutral statement and I is a conjunction of indexical statements. Then a scenario characterized by D&I will correspond to a centered world w iff D is true of w and iff I is true (in the obvious sense) of the individuals at the center of W.)

When S is a strong necessity, by contrast, there will be a scenario verifying $\neg S$, but this scenario will correspond to no metaphysically possible world. (When S is a positive strong necessity,

there will be a positive scenario verifying $\neg S$; when S is a negative strong necessity, there will be a negative scenario verifying $\neg S$.) For example, given the theist view outlined above, there will be a (negative and positive) scenario verifying 'There is no omniscient being', involving some maximally detailed hypothesis under which there is no such being. But on this view, there is no centered world that corresponds to this scenario, and there is no centered world that itself verifies 'There is no omniscient being'. We might put this intuitively by saying that on this view, the space of (centered) metaphysically possible worlds is *smaller* than the space of epistemically possible scenarios, at least in the relevant respect. On this view, there are scenarios that correspond to no world.

To bring this back to the mind–body case: take the paradigmatic type-B materialist who holds that premise (1) is true, premise (2) is false, and materialism is true. On this view, the material conditional $P \supset Q$ (which is itself the negation of $P\&\neg Q$) is a strong necessity. The truth of materialism implies that it is 2-necessary, the truth of (1) implies that it is a posteriori and its negation is 1-conceivable, but the falsity of (2) implies that its negation is 1-necessary. On this view, there will be a *scenario* verifying $P\&\neg Q$, including various specific zombie scenarios. But these scenarios will correspond to no metaphysically possible world.

Note that the analog of CP with scenarios instead of worlds is close to trivial: if S is conceivable in the relevant sense, there will automatically be a scenario verifying S (at least if the notions of a scenario and of verification are unproblematic). Even a type-B materialist and a believer in strong necessities can accept that principle. They must simply deny that all scenarios correspond to worlds. So CP might be seen as equivalent to the thesis that for every scenario, there is a corresponding world.

(In Chalmers 2004a, the Core Thesis of epistemic two-dimensional semantics says that S is a priori iff S has a necessary 1-intension. If 1-intensions are defined over scenarios, then the resulting Weak Core Thesis is independent of CP, makes no claim about metaphysical possibility, and need not be disputed by believers in strong necessities. If 1-intensions are defined over centered worlds, then the resulting Strong Core Thesis is equivalent to CP and must be disputed by believers in strong necessities. The Strong Core Thesis can also be understood as the thesis of Metaphysical Plenitude: that every negatively conceivable statement is verified by some centered world. So the nexus of the debate over premise (2) might be phrased in terms of either CP, the Strong Core Thesis, or the thesis of Metaphysical Plenitude.)

My view is that for every scenario there is a corresponding world, and that there are no strong necessities. In *The Conscious Mind*, I gave the following reasons for this: (i) strong necessities

cannot be supported by analogy with other a posteriori necessities; (ii) they involve a far more radical sort of a posteriori necessity than Kripke's, requiring a distinction between conceptual and metaphysical possibility at the level of worlds; (iii) they lead to an *ad hoc* proliferation of modalities; (iv) they raise deep questions of coherence; (v) strong necessities will be brute and inexplicable; and (vi) the only motivation to posit a strong necessity in the mind–body case is the desire to save materialism. I still accept most of these reasons, but there is more to say.

In the last decade or so, numerous objections to CP have been proposed. These objections fall into a number of classes. The first, and largest, involves attempts at exhibiting clear cases of strong necessities. The second involves attempts at explaining how there might be strong necessities in the phenomenal case (if not elsewhere), by analyzing the nature of phenomenal concepts. The third involves general philosophical objections. I will take these in turn.

8 Are There Strong Necessities? I: Examples.

I will discuss a number of putative counterexamples to CP in turn. The first five involve a posteriori identities. The next four involve are tied to issues about the a priori and a priori entailment, and involve potential challenges to CP-. The last five, suggested especially by Yablo (1999, 2002), involve challenges to CP- and CP+.

(1) *Kripke cases*: It is occasionally proposed that some Kripkean a posteriori necessities are in fact strong necessities. In particular, it is sometimes proposed that coextensive names such as 'Cicero' and 'Tully' may have the same primary intension as well as the same secondary intension. If so, 'Cicero is Tully' is a strong necessity (as it is clearly a posteriori).

In response: When we entertain the hypothesis that Cicero is not Tully, this hypothesis corresponds to specific scenarios that we can elaborate. In particular, the relevant scenarios may involve the hypothesis that the causal chains associated with names 'Cicero' and 'Tully' pick out different historical individuals. A scenario like this certainly corresponds to a centered metaphysically possible world: there are certainly worlds where the causal chains associated with these words functions in this way. And it seems clear that if we discovered that our world were such a world, we would reject the hypothesis that Cicero is Tully. So such a world seems to verify 'Cicero is not Tully' (although it may not satisfy 'Cicero is not Tully'). Worlds like this can be found for any Kripkean a posteriori necessity S (as Kripke himself pointed out, in effect), and such worlds will always verify $\neg S$.

One might resist the claim that the world in question verifies S by rejecting the claim that

there is an *a priori* entailment (for the speaker in question) from a description of the world in question to S. In response, I think that the sort of considerations in Chalmers and Jackson (2001) strongly suggest that these entailments are a priori at least in principle. But even if one rejects this claim, there is clearly remains *some* distinctive epistemic relation between the world in question and 'Cicero is not Tully': in particular, it remains the case that if one accepts (hypothetically) that the actual world is qualitatively just like the world in question, and reflects on this hypothesis, then one will reject the claim that Cicero is not Tully. (Note that this is quite unlike the situation that the theist thinks obtains in the God case, where there is no world that stands in this inferential relation to 'There is no God'.) So if one rejects a priori entailments, one can use this sort of inferential relation to define primary intensions, and ordinary Kripkean necessities will always have a primary intension that is false at some world.

(2) *Essential modes of presentation*. Block (??) and Tye(??) have suggested that one might get counterexamples to CP via identities in which each expression picks out an identity via an essential property. For example, let Q be a quantum-mechanical description that happens to pick out $\mathrm{H_2O}$. Then 'Q=$\mathrm{H_2O}$' is a posteriori (since the quantum-mechanical and chemical descriptions are not conceptually connected), but each picks out the referent via an essential property (which corresponds to the primary intension), so 'Q=$\mathrm{H_2O}$' will be 2-impossible. Kallestrup (forthcoming) also argues that a posteriori microphysical identities such as 'Q=$\mathrm{H_2O}$' will be strong necessities, on the grounds that microphysical terms have identical primary and secondary intensions.

In response: It is arguable that microphysical and chemical expressions pick out their referents by *contingent* properties, for reasons described earlier. Further, an entity can have more than one essential property, so that if a substance is picked out by *distinct* essential properties, the corresponding primary intensions will differ and the sentence will not be 1-possible. Finally, it is pretty clear that the scenario one has in mind when one conceives of the falsity of 'Q=$\mathrm{H_2O}$' (say, a world where something with the structure of Q yields quite different chemical-level properties due to different quantum-mechanical laws) will itself correspond to a possible world W that verifies 'Q is not $\mathrm{H_2O}$'. So unless one gives independent reason to think that there is no such world w (perhaps because one holds the strong view of the necessity of laws of nature, discussed under objection (12) below), there is little reason to deny the natural view that 'Q' and '$\mathrm{H_2O}$' have distinct primary intensions. If so, there is no counterexample here.

(3) *Distinct homophonous expressions*. Block (2006) has suggested a case analogous to the 'Cicero=Tully' case above in which the same orthographic expression is used on each side. E.g., perhaps one has acquired the term 'chat' on two separate occasions for two separate species. One

may then lose any associated information, but still recognize that they are distinct expressions, so that 'chat=chat' will not be a priori. But let us say, the two expressions 'chat' pick out the same referent. Then 'chat=chat' will be necessary, but both sides will have the same primary intension. At least, the explanation above, suggesting a primary intension along the lines of "the thing causally connected to my use of 'chat"', suggests that the sentence will be 1-necessary.

In response: Even here, there will be a centered world corresponding to the scenario one has in mind when one entertains the falsity of the hypothesis one expresses with 'chat=chat'. Such a centered world involves distinct causal chains from each of the two uses of 'chat' to distinct entities. Here the two uses can at least be distinguished *indexically*, with a primary intension along the lines of "the thing causally connected to *this* use of 'chat"'. (If necessary, these indexicals can be grounded in direct indexical reference to token mental states, along the lines below.)

(4) *Demonstratives*: Schiffer (2004) suggests a worry about identities involving perceptual demonstratives, such as 'that=that'. In most such cases, there is no problem: the conceivable falsity of such a claim corresponds to a scenario in which the item causing a certain sort of experience in one differs from the item causing a different sort of experience, and this scenario will correspond to a centered world, such that if one discovered this world to be actual then one would reject 'that=that'. But there are cases in which the relevant experiences are indistinguishable, such as Austin's "Two Tubes" case (1989) in which one simply sees two red dots in a symmetrical visual field, or a case where one sees many dots and has no way to pick out any one of them by description. Then the model above involving different sorts of experience will not apply.

In response: Again one will have to invoke indexicals here, so that the primary intension of each 'that' will be roughly equivalent to that of 'the cause of *this* experience', for different phenomenal indexicals 'this'. In this case, the phenomenal indexical must be considered a primitive indexical akin to 'I' and 'now'. Formally, evaluating the primary intension of such an indexical requires a centered world where the center includes not just a subject and a time, but certain marked experience tokens, to which the distinct phenomenal indexicals are linked. Then 'that=that' will be false at centered worlds where the two relevant experience tokens marked at the center are caused by different objects. For more on this, see Chalmers (2003b).

(5) *Dancing qualia*: Hawthorne (2006) suggests a case involving phenomenal concepts, based on the dancing qualia case in *The Conscious Mind*. Say that Fred starts with identical red experiences in the left and right half of his visual field, forms concepts *R1* and *R2* (pure phenomenal concepts, in the language of Chalmers 2003b) of the relevant phenomenal property (phenomenal redness) and judges *R1=R2*. After this the experiences in the right half of his visual field "dance"

between red and blue every minute, while they continue to play the same functional role (in *The Conscious Mind* I argued that this is naturally impossible but allowed that it is logically possible). Fred has been told that this may happen, but when a change happens he is unable to notice it. After four minutes his right-field experience is back to red, and he once again entertains the thought *R1=R2*. But he is unable to know that this is true through any amount of a priori reasoning, so *R1 is not R2* is at least negatively conceivable. But both *R1* and *R2* have the same primary intension, so *R1=R2* is 1-necessary. If so, this is a counterexample to CP-.

In response: This should be treated analogously to a case in which Fred's cognitive processes are tampered with so that they first judge *7+5=12* (with normal processes), then *7+5=14* with abnormal cognition, then back again. If Fred knows this is happening, then when he entertains *7+5=12* he will be unable to know that it is happening through any amount of a priori reasoning. Nevertheless, his thought is still a priori. This is simply a situation in which tampering with cognitive processes renders him less than ideal, so that he is unable to know all a priori knowable truths. (Or, in a closely related version, it is a situation where misleading empirical evidence undercuts his ability to know these.) The dancing qualia case above should be treated the same way.

(6) *Ordinary macroscopic truths* Another class of examples comes from the suggestion that consciousness is not alone in its failure to be a priori entailed by microphysical truths. Block and Stalnaker (1999) have suggested that many ordinary macroscopic truths, such as "water boils at 100 degrees" (or W), are not a priori entailed in this way. If so, then $P\&\neg W$ is at least negatively conceivable. But P&W is not possible (at least if water and its properties supervene on the microphysical), and it is not clearly 1-possible, either. If it is not, then $P\&\neg W$ is a counterexample to CP-.

In response: Chalmers and Jackson (2001) argue that these macroscopic Ctruths are a priori entailed by P, or at least by $PQTI$ (say what this is, here if not before). Of course $P\&\neg Q$, $P\&\neg T$, and $P\&\neg I$ may be negatively conceivable, but the latter two are clearly also 1-possible, while the former is precisely the main topic of dispute. So any failures of entailment by P associated with the failure of entailment to Q, T, or I give no further support to the existence of strong necessities. Such support would require at least a truth M such that $PQTI$ does not a priori entail M, and such that $PQTI \supset M$ is 2-necessary and 1-necessary. The arguments in Chalmers and Jackson 2001 suggest that at no ordinary macroscopic truths M are like this.

In any case, even if one rejects the a priori entailment thesis here, these cases will yield at best exceptions to CP- (i.e. negative strong necessities), not exceptions to CP+ (i.e. positive

strong necessities). Even if w above is like Q in that $P\&\neg W$ and even $PQTI\&\neg W$ is negatively conceivable, $PQTI\&\neg W$ is not positively conceivable: "water-zombies" are not positively conceivable in the way that zombies are positively conceivable. So this sort of case leaves CP+ unthreatened.

(7) *Unknowable mathematical truths*. Perhaps the most challenging cases for CP- are mathematical truths M such that M is true (and so necessarily true and 1-necessarily true), but not knowable, and so not knowable a priori. If there are such cases, M is a negative strong necessity (though not a positive strong necessity, as $\neg M$ is not positively conceivable). Here one might appeal to unprovable true mathematical sentences, such as those whose existences is entailed by Gödel's theorem.

In response: Unprovability in a given system does not entail non-apriority. For example, the consistency of Peano Arithmetic is not provable in Peano Arithmetic, but is still plausibly knowable a priori. One can make the case (see Chalmers 2002) that all true statements of arithmetic are knowable a priori at least under an idealization (i.e., our failure to know them a priori is due to certain limitations of our cognitive systems). One might worry about higher set theory, such as the Continuum Hypothesis, but here it is far from clear that such sentence are determinately true or false, and it is also far from clear that they are not knowable a priori under an idealization. So although these cases provide an interesting challenge to CP-, they do not provide clear counterexamples to CP-. And as in the previous case, there is no challenge to CP+ here.

(8) *Inscrutable truths*. A related class of potential counterexamples to CP- includes special sorts of true sentences M such that according to certain philosophical views, M is not a priori entailed by any underlying qualitative truths, but such that it is necessitated by such truths. In Chalmers (2002) I call these *inscrutable truths*. Such sentences might include vague sentences such as 'John is bald', under some versions of the epistemic theory of vagueness, according to which such sentences can be true but unknowable even in principle on the basis of underlying qualitative information. Others might include certain moral or metaphysical truths, on views according to which there are true such sentences with no a priori link to underlying qualitative truths. In Chalmers (2002) I argue that there are no compelling counterexamples here, as the views in question should all be rejected. In any case, these cases do not typically go along with the positive conceivability of $PQTI \supset \neg M$, and so at best challenge CP-, not CP+.

(9) *The deeply contingent a priori*. Hawthorne (2004) suggests that some truths may be a priori even though they are false in some possible worlds considered as actual. For example, let $E \supset H$ be a conditional from one's total evidence to a hypothesis it supports. Then there are some

grounds for holding $E \supset H$ to be a priori (since the inference from E to H transmits justification and arguably cannot itself be a posteriori). But $E \supset H$ it is certainly not 1-necessary: $E\&\neg H$ is verified by many worlds where the evidence is misleading. If so, $E\&\neg H$ is ruled out a priori (so it counts as inconceivable by the earlier definition), but it is 1-possible. Strictly speaking this sort of case does not challenge CP-, but rather challenges the associated "right-to-left" principle. This is still a worry for the framework, however, and addressing it helps to clarify an important issue.

In response: There may be a sense in which $E \supset H$ is a priori here, in that it has some justification independently of experience. But there is a sense in which it is not, or at least in which its justification is much weaker than that of many other a priori truths. Intuitively, one cannot *conclusively* rule out the possibility that $E\&\neg H$, whereas one can conclusively rule out the hypothesis that 2+2=5 for example. Conclusive justification is the sort that one gets from proof and analysis, rather than from abduction and induction, for example. In the uses of the a priori in the current context, it is always the conclusive a priori that is relevant. By this standard, $E\&\neg H$ is not ruled out a priori. This allows one to accommodate the clear intuition that $E\&\neg H$ is not just 1-possible but 1-conceivable: it is always conclusive apriority that is relevant to matters of conceivability.

(10) *Disquotational truths*. Yablo (2002) suggests that certain disquotational truths, such as "'sister' means sister", are a priori but are not 1-necessary. It could turn out that 'sister' does not mean sister, in worlds where the term is used differently, and such worlds verify "'sister' does not mean sister".

In response: I think the treatment of this case differs depending on whether one sees 'sister' picking out its referent orthographically (so just a certain orthographic form is required) or semantically (so that a certain meaning is required). If "'sister"' is understood orthographically, then "'sister' means sister" is not a priori: it is substantive a posteriori knowledge that the orthographic item means anything at all. If "'sister"' is understood semantically, then "'sister' means sister" may be a priori, but it is also 1-necessary: cases where the orthographic item is used with a different meaning will be irrelevant to the primary intension of "'sister"'.

(11) *Response-enabled concepts*. Yablo (2002) argues that certain geometric terms such as 'oval' may pick out certain shapes according to the way they trigger certain responses, rather than through a substantive independent grasp. In the actual world 'oval' picks out cassinis (where 'cassini' is defined mathematically), but 'ovals are cassinis' is not knowable a priori. But 'ovals are cassinis' is nevertheless necessary and 1-necessary (a world where cassinis cause oval experiences is not a world where it would have turned out that ovals are cassinis, according to Yablo), so it is

a strong necessity.

In response: I think that the actual term 'oval' may embody a geometric understanding, so that 'ovals are cassinis' (or something like it) is a priori. But there might be other terms that could work in Yablo's response-enabled way, though, so I will go along with his treatment of 'oval' as one of them. Under this assumption, we should say that 'ovals are cassinis' is not 1-necessary. If w is a world where Hs cause oval experiences, then (under this model of 'oval') it is a priori that if w is actual, so that Hs cause oval experiences, then ovals are Hs. (What matters for evaluating a primary intension is the a priori connection, not the "turns out" claim above.)

(12) *Laws of nature*. According to some philosophical views, laws of nature are not just naturally necessary but are metaphysically necessary. Shoemaker (1999) holds such a view, and suggests that it may provide a counterexample to CP.

In response: On some versions of this view, laws of nature follow a version of the Kripkean model. That is, if mass in the actual world obeys certain laws, then nothing in any counterfactual world counts as mass unless it obeys exactly those laws, so any law involving mass will be necessary. This might hold because of the semantics of mass (which require that a counterfactual property have the same nomic role as actual mass in order to qualify as the referent of 'mass'), or it might hold because of the metaphysics of mass (according to which properties such as mass have their nomic profile essentially, as on Shoemaker's view). On these models, we need not deny that there are worlds that correspond to the scenario we conceive when we conceive that mass obeys different laws: it is just that such worlds will contain "schmass" rather than mass. I think that it is implausible that the modal profile of 'mass' and/or the essential properties of mass are this precise (see Fine 2002 and Sidelle 2002 for arguments), but in any case, this model does not provide a counterexample to CP. In this case, a schmass world may verify the hypothesis that the relevant law of nature is false, so laws of nature are not strong necessities.

To yield strong necessities, this sort of view must hold that not only are there no worlds where mass obeys different laws, but there are also no related worlds where something else, "schmass", obeys those different laws. Here, the relevant sort of view is one according to which the fundamental properties and laws of all worlds are the fundamental properties and laws of our world (and on which these laws are not knowable a priori). In effect, this restricts the space of metaphysically possible worlds to the space of naturally possible worlds. If this view is correct, then a fundamental law will be a strong necessity: there will be no world corresponding to the scenario that we conceive when we conceive that is false.

I think that there are no good reasons to accept this extremely strong view of laws of nature,

and that there are good reasons to reject it. The best reasons to take the hypothesis that laws of nature are necessary seriously come from the Kripke and Shoemaker models above. But nothing in these models supports the strong view, or yields a strong necessity. Rather, the CP thesis can itself be taken as reason to reject the view.

(13) *A necessary God*. I have already noted that the existence of strong necessities is entailed by a theist view on which an omniscient being (or an omnipotent being, or a perfect being) exists necessarily but on which the existence of such a being is not knowable a priori. If we say that a god is by definition an omniscient being (or a perfect being, or whatever), then 'A god exists' will be a strong necessity.

In response: I think that theist views of this sort are to be rejected. If the existence of such a god is knowable a priori, then it may exist necessarily. But if it is not, then one should conclude that such a being exists at best contingently. I cannot go over the arguments for believing in a necessary god here, but they all rest on highly contentious premises, and once again, CP itself provides an argument against these views. The best way to defend the existence of a necessary god is to argue that a world without such a being is not even conceivable.

Even if one believes in that the existence of a god provides a strong necessity, it is not clear that this sort of strong necessity undermines the case against materialism. The debate over materialism uses necessity as a criterion of ontological distinctness: the question of whether physical truths necessitate phenomenal truths is relevant precisely insofar as it bears on the question of whether the phenomenal involves nothing ontologically "over and above" the physical. But a variety of necessity in which the existence of a god is necessary will not be well-suited to this role. On such a view, the existence of a god will be necessitated by physical truths, but such a god will presumably nevertheless be ontologically nonphysical. So if the only strong necessities are strong necessities of this sort, connecting ontologically distinct existences, their existence is no help to physicalism. Under this assumption, then if $P\&\neg Q$ is conceivable, Q will be something over and above the physical, either because it is necessitated by the physical or because it is tied to the physical only by this sort of strong necessity.

Something similar applies to views on which laws of nature are strong necessities. Even on such views, laws presumably connect ontologically distinct properties: if there is a fundamental law connecting properties A and B, this will not ground any sort of ontological reduction of one property to the other. Indeed, if this view is correct, then a dualist view with fundamental phenomenal properties and fundamental laws connecting them to the physical will itself be a view on which the phenomenal is necessitated by the physical. So again, strong necessities of this sort are

no help to the physicalist.

(14) *Metamodal claims*. Yablo (1999) adapts the God case to provide an intriguing argument against CP. According to Yablo, it is at least *conceivable* that there is a necessarily existing god. It is also conceivable that there is no necessarily existing god. So, if G is 'It is necessary that there is an omniscient being', then both G and $\neg G$ are conceivable. If so, then by CP both G and $\neg G$ are 1-possible. There appears to be no relevant distinction between the primary and secondary intensions of the expressions involved, so it follows that G and $\neg G$ are 2-possible, or (metaphysically) possible simpliciter. So it is possible that it is necessary that there is an omniscient being, and it is possible that it is not necessary that there is an omniscient being. But this is a contradiction, at least given S5 as the logic of the metaphysical modality. If it is possible that S is necessary, then S is necessary, so it is not possible that S is not necessary.

In response: One could respond by denying S5, or by finding relevant two-dimensional structure, but I think these moves are unpromising. One could also respond, more promisingly, by making the observation about the ontological relevance of this sort of necessity above. But I think it is best to deny that it is conceivable that there is a necessarily existing god, at least in the relevant sense of conceivability. Perhaps it is prima facie negatively conceivable that there is such a being, in that we cannot obviously rule it out a priori, but I do not think it is conceivable in any stronger sense. I can certainly 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. Once one accepts that it is conceivable that there is no god (and this seems like a much stronger intuition, at least to me), this has a strong tendency to undermine the coherence of the hypothesis that a god exists necessarily.

The problematic issues here arise because of the double modality: we are conceiving not just of nonmodal qualitative features of worlds, 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, despite the conceivability of a godless world. That is, we have to conceive that CP is itself false. This is what does all the work in the example: if it is conceivable that CP is false, then (by CP!) it is possible that CP is false. CP is surely necessarily true if it is true at all, so it follows from the possible falsity of CP that CP is false.

Another way to respond 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 (CP would then apply to worlds considered nonmodally, but not to "cosmoses" of possible worlds). But I prefer to hold onto the stronger thesis, by denying that it is conceivable that CP is false. I hold that CP is a priori, although highly nontrivial, like many theses in philosophy. In fact I will sketch an a priori argument for CP later in this paper. If this is correct, then CP is not conceivably false on ideal rational reflection, and it is not ideally conceivable that a necessarily existing god exists.

(15) *The conceivability of materialism*. A closely related metamodal argument (Marton 1998; Sturgeon 2000) that is specific to the mind–body domain proceeds as follows. (i) It is at least conceivable that materialism is true about consciousness. So (ii) it is conceivable that $P \supset Q$ is necessary. By CP (and setting aside two-dimensional structure), it follows that (iii) it is possible that $P \supset Q$ is necessary. But from this it follows (using S5) that (iv) $P \supset Q$ is necessary. Using CP and S5, one can equally infer from the fact that (v) it is conceivable that $P \supset Q$ is not necessary to the conclusion that (iv) $P \supset Q$ is not necessary. But (iv) and (vi) are contradictory. So one should reject CP.

My response here parallels the response in the god case. It may be prima facie negatively conceivable that materialism is true about consciousness, but it is not obviously conceivable in any stronger sense. Many people have noted that it is very hard to imagine that consciousness is a physical process. I do not think this unimaginability is so obvious that it should be used as a *premise* in an argument against materialism, but likewise, the imaginability claim cannot be used as a premise either. And if I am right that CP is a priori, then there is an a priori argument that $P \supset Q$ is not necessary, so that it will not even be ideally negatively conceivable that $P \supset Q$ is necessary.

Overall. We have seen that there have been many attempts at providing counterexamples to CP, but none of these provides a clear counterexamples. In most cases, I think there are reasonably straightforward independent grounds for rejecting the claim that the cases in question provide strong necessities. Perhaps the most serious challenges come from mathematical cases such as the Continuum Hypothesis and from metamodal cases. These are the cases where, in advance of commitment to CP, independent reasoning does not clearly settle whether or not strong necessities are involved. Still, these are cases where the initial situation is unclear, rather than cases where there is a clear counterexample. If one can make the case for CP independently, then these cases are not too much of a threat.

It is also worth noting that cases such as these seem to work best as challenges to CP- rather than to CP+, so that CP+, which is all that is required for the argument against materialism,

is relatively unthreatened. We have also seen at various points along the way that even if one takes certain cases to involve strong necessities, the existence of such strong necessities will still be compatible with modified versions of CP (say, a version involving ontological necessities in the law/god cases, or a version involving nonmodal sentences in the metamodal cases) that will be strong enough for the anti-materialist argument go through. So where the consideration of counterexamples is concerned, the anti-materialist seems to be on reasonably strong ground.

9 Are there strong necessities? II: Explanations

In the previous section, I considered attempts on behalf of type-B materialism to establish that there are strong necessities outside the mind–body domain, and argued that these attempts fail. Other type-B materialists accept that there are no strong necessities outside the mind–body domain, but argue that there are nevertheless strong necessities inside the mind–body domain. To avoid postulating an ad hoc exception, such a view must provide an *explanation* of why the mind–body domain is exceptional.

By far the most common way to explain such exceptions is to appeal to the special nature of phenomenal concepts, the concepts that we use to think about consciousness. Proponents argue that because of this nature, we should expect that there are strong necessities involving consciousness, reflecting not an ontological gap in nature but rather some a merely epistemic gap in our cognitive processes.

I think that this strategy is the most interesting and attractive strategy for the defense of type-B materialism. However, I do not think that it can work. In "Phenomenal Concepts and the Explanatory Gap" (Chalmers 2006a), I argue that no such strategy can succeed in supporting type-B materialism against the arguments considered here. Here, I will briefly consider two versions of the strategy, and will offer some different arguments against them (drawn from Chalmers 1999).

(1) *Independent cognitive processes*. Hill (1997) and Hill and McLaughlin 999) offer a psychological explanation of why we can conceive of zombies, in terms of the independent cognitive processes by which we conceive of physical processes and by which we conceive of experiences. On their view, these processes involve independent faculties, and the mental representations involved have independent cognitive roles, so it is to be expected that there is no conceptual connection between them. But this independence of the cognitive processes can itself be given a physical explanation.

In response: the crucial question is whether a psychological explanation of the independence

of physical and phenomenal concepts suffices to explain the existence 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 involved in conceiving of color and shape. One can give a psychological explanation of why we can conceive of five-horned animals, 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.

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 the possibility of such states-of-affairs. This is what Kripke (1980) does, in "explaining away" the intuition that heat might not have been molecular motion: he explains why and how the conceivability of this state of affairs if compatible with its impossibility. Hill (1997) likens his strategy to Kripke's, but it discharges only half of the burden: it explains why zombies might be conceivable, but it does nothing to explain why and how this conceivability coexists with the impossibility of zombies. Without such an explanation, a strategy like Hill and McLaughlin's could be used to "explain away" any conceivability intuition at all. Presumably, there will always be a psychological explanation of the processes involved in a modal intuition. But one presumably should not infer that they are unreliable. If one did, then one might likewise find a psychological explanation of our mathematical beliefs, and infer that these beliefs are no guide to mathematical truth.

Hill and McLaughlin come close to addressing this issue by 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". 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. As it stands, Hill's and McLaughlin's strategy does not provide such an explanation.

(2) *Recognitional concepts*. Loar (1997; 1999) attempts to fill this gap, by explaining how physical and phenomenal concepts can be conceptually distinct but nevertheless pick out the same property, and without distinct properties serving as modes of presentation. If one could do this, one would thereby explain why identities involving such concepts are strong necessities.

Loar appeals to two theses about phenomenal concepts. First, he appeals to the thesis that they are recognitional concepts, picking out their referent through a direct process of recognition.

Second, he appeals to the thesis that they have the same primary and secondary intensions. (Loar puts this in his own terminology, by saying that phenomenal concepts have the same property as both referent and reference-fixer, but for ease of discussion I will translate into my terminology.) The significance of the first thesis is that a recognitional concept ("*that* sort of cactus") may refer to the same property as a physical-theoretical concepts, even though the two concepts are cognitively distinct. On its own, this is not enough to yield a strong necessity: recognitional concepts and physical-theoretical concepts typically have quite different primary intensions, partly because the recognitional concept picks out its referent under a contingent mode of presentation (e.g., a cactus might be picked out by relying on the sort of experience it produces). So Loar adds the second thesis, holding that phenomenal concepts are unique among recognitional concepts in having the same primary and secondary intensions. Loar also holds that theoretical physical concepts have coinciding primary and secondary intensions, an assumption that I will go along with here for the purpose of simplifying discussion. (As usual, relaxing the assumption only opens up the possibility of Russellian monism.)

From the first thesis, Loar infers that phenomenal concepts may corefer with physical concepts even though they are cognitively distinct (in such a way that zombies are conceivable). Both concepts are rigid, so they will have coinciding secondary intensions. From the second thesis (combined with the same claim about physical concepts), he infers, in effect, that the two concepts have coinciding primary intensions. So an identity involving the concepts will be 1-necessary, despite being a posteriori and counterconceivable. Such an identity will be a strong necessity.

The problem with this is straightforward. The truth of the second thesis undercuts Loar's inference from the first thesis. It may be true that recognitional concepts are cognitively distinct from theoretical concepts, and it may be true that they *often* corefer with those concepts. But the cases in which they do are all cases where the recognitional concepts have nontrivial two-dimensional structure, typically pick out out their referent as the cause of a certain sort of experience, or by some similar contingent mode of presentation, so that they have distinct primary and second intensions. Furthermore, this two-dimensional structure provides a clear *explanation* of how and why the recognitional concept coqrefers with a theoretical concept. 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.

Loar's two theses, if accepted, should lead one to accept that phenomenal concepts and physical concepts (i) are cognitively distinct, and (ii) both have the same primary and secondary intensions. But 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 recognitionality is doing some extra work here (thus distinguishing this case from other cases involving nonrecognitional concepts), but it merely supports (i), and supports the possibility of coreference when (i) is true but (ii) is false. So nothing here supports the possibility of coreference when (i) and (ii) are both true. Rather, given the truth of (i) and (ii), one should more naturally infer that the referents of the concepts differ.

Loar suggests that his two theses are themselves neutral between physicalism and dualism, and so can equally be conjoined with the claim that phenomenal concepts pick out physical properties as with the claim that they pick out nonphysical properties. But this is far from obvious: the reasoning above suggests that the two theses support the claim that phenomenal concepts pick out nonphysical properties. Perhaps Loar could build in the further claim that phenomenal concepts pick out physical properties as part of his explanation, but then this claim will be doing all the work. Loar's opponent will deny the possibility of this claim, so Loar still needs an explanation of how phenomenal concepts *could* pick out physical properties, given the situation above. As things stand, the model above will then require that physical properties have their phenomenal mode of presentation non-contingently. But this means that the explanation is building in a necessary connection between physical and phenomenal properties from the start, and so is assuming strong necessities in order to explain strong necessities.

It appears that neither Hill & McLaughlin's nor Loar's account can explain the existence of strong psychophysical necessities. Instead, both accounts need to assume such necessities at a key point, as a kind of primitive. One might think that some more refined account will avoid this problem, but I think that the problem is inevitable. For an in-depth discussion of problems that afflict all appeals to phenomenal concepts to support physicalism, see Chalmers (2006a).

10 Modal Rationalism

So far, I have argued that there are no clear counterexamples to the conceivability-possibility thesis, and no good explanations of how it could be false. Most type-B materialists support their view either by potential counterexamples to conceivability-possibility theses or by potential explanations of its falsity, so I think this removes the central plank of support from the view. Still, one might wonder why CP *has* to be true. Why couldn't there be strong necessities, so that some scenarios correspond to no metaphysically possible world?

The first thing one can say at this point is that in the absence of an explanation, these strong

necessities will be brute and inexplicable. Epistemically, they must be taken as primitive, in the same way that we take fundamental laws of nature as primitive. We might consider them to be "fundamental laws of metaphysics", or as laws grounded in related fundamental laws involving properties, essences, and the like. I think that there is good reason to doubt that there are a posteriori fundamental laws of this sort. Rather, where there are a posteriori fundamental laws, these should always be taken as laws of nature, and therefore as metaphysically contingent.

I think that the most fundamental reason for rejecting strong necessities comes from a conceptual analysis of modal notions, and an analysis of the reasons that lead us to believe in modality in the first place. This analysis is itself an extended project, but here I will give the sketch of an argument.

The argument involves locating the roots of our modal concepts in the rational domain. When one looks at the purposes to which modality is put (e.g. in the first chapter of Lewis 1986), it is striking that many of these purposes are tied closely to the rational and the psychological: analyzing the contents of thoughts and the semantics of language, giving an account of counterfactual thought, and analyzing rational inference. It can be argued that for a concept of possibility and necessity to be truly useful in analyzing these domains, it must be a rational modal concept, tied constitutively to consistency, rational inference, or conceivability.

It is not difficult to argue that even if not all conceivable worlds are metaphysically possible worlds, we *need* a modal concept tied to rational consistency, apriority, or conceivability to best analyze the phenomena in question. We might call this modality the *logical* modality. We can say that S is logically necessary when S is a priori, and we can define corresponding notions of possibility and entailment. We can then argue that a space of logically possible *worlds* will be vital for many of the explanatory needs for which possible worlds are needed in the first place.

To see this, let us pretend for a moment that all worlds with laws of nature that differ from ours are metaphysically impossible. Even so, it will still be tremendously useful to appeal to a wider space of logically possible worlds (or world-like entities) with different laws, to help analyze and explain the hypotheses and inferences of a scientist investigating the laws of nature. Such a scientist will be considering all sorts of rationally coherent possibilities involving different laws; she will make conditional claims and engage in counterfactual thinking about these possibilities; and she may have terms and concepts that are co-extensive at all worlds with our laws, but that intuitively differ in meaning because they come apart at worlds with different laws. To analyze these phenomena, the wider space of worlds is needed to play the role that possible worlds usually play.

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.

Further, there is no bar to a space of such worlds. If one does not want simply to postulate them, one can easily construct them in an 'ersatz' way, perhaps using equivalence classes of sets of semantically neutral sentences. (If one has qualms about using the term 'world' for these entities, nothing turns on the word: one can equally call them 'scenarios', or some such, instead.) One can then introduce means of semantically evaluating expressions at these worlds, by considering these worlds as actual (considering what to infer from the hypothesis that the world actually obtains) or as counterfactual (considering what would have been the case had the world obtained). There are a few complications: one may want to add centers to the worlds for consideration as actual, and perhaps to allow haecceitistic differences between worlds for consideration as counterfactual, depending on one's philosophical views. But this sort of construction is reasonably straightforward (a related construction is given in Chalmers forthcoming).

The two sorts of evaluation over these worlds corresponds to two sorts of necessity of sentences. We might say that S is *epistemically necessary* if it is true at all logically possible worlds considered as actual, and that S is *subjunctively necessary* if it is true at all logically possible worlds considered as counterfactual. Truths involving the epistemic modality are in general a priori, while truths involving the subjunctive modality may be partly grounded in nonmodal truths about the actual world. When a sentence S is subjunctively but not epistemically necessary, it will be an a posteriori necessity. If property terms A and B are such that 'Something has A iff it has B' is subjunctively but not epistemically necessary, then A and B will express distinct concepts while referring to the same property. (At least this is so for on a coarse-grained individuation of properties; for familiar reasons, fine-grained individuation requires going beyond possible worlds.)

On this analysis, we have two modal concepts, epistemic and subjunctive necessity. But importantly, the two concepts are constitutively linked, to each other and to the rational domain. We might see both as varieties of (ideal) conceivability: epistemic possibility involves what conceivably *might be* the case, and subjunctive possibility involves what conceivably *might have been* the

case. Both are ultimately grounded in what it is rationally coherent to suppose.

These logically possible worlds and semantic evaluation over them yield a modal space that is useful for all sorts of purposes. In fact, these worlds will be useful for precisely the purposes for which possible worlds are needed in the first place. This modal space is perfectly suited to analyze such rational and psychological matters as counterfactual thought, rational inference, and the contents of thought and language. And through two-dimensional semantic evaluation, we have seen that it can yield such "metaphysical" modal phenomena as the concept/property distinction, a posteriori necessities, and so on.

At this point, an opponent of the CP thesis might allow that the space of logically possible worlds is coherent and useful in its own right. But they may well suggest that we also have good reason for believing in a separate space of metaphysically possible worlds. Presumably this space will be narrower than the first space. While the space of logically possible worlds includes worlds with zombies, worlds with different laws, worlds without gods, and so on, the space of metaphysically possible worlds may exclude some or all of these.

In response, one can first argue that the space of logically possible worlds, along with the two sorts of evaluation, suffices to account for all modal phenomena that we have reason to believe in and that we might invoke possible worlds to explain. Certainly it accounts for the data about counterfactuals, inference, and content at least as well as the ordinary space of possible worlds does. And it also accounts for the Kripkean data, concerning intuitions about a posteriori necessities and the like. While it does not account for "data" about the necessity of laws or of gods, these cannot be considered clear and untendentious data. Rather they are theoretical claims whose status is up in the air, to be settled by the best account of modality.

One can then argue that there is no good reason to postulate a separate space of metaphysically possible worlds. There is no clear explanatory work left for such a space to do. The space of logically possible worlds, which we have independent reason to postulate, explains all the untendentious modal data. It is not that there is no such thing as metaphysical necessity. Rather, metaphysical necessity is simply subjunctive necessity over the space of logically possible worlds. To introduce a further primitive, restricting the space of worlds, is to introduce an unnecessary wheel.

To resist, an opponent might do one of four things. First, they might argue that logically possible worlds are problematic in their own right: something about their construction or the associated sorts of semantic evaluation is ill-defined or incoherent. Second, they might argue that logically possible worlds cannot explain the phenomena: there are untendentious modal data that logically

possible worlds cannot explain, but that a separate space of metaphysically possible worlds can explain. Third, they might argue that although logically possible worlds can accommodate the untendentious phenomena, there is good theoretical reason to believe in a distinct space of metaphysically possible worlds all the same. Fourth, they might argue that we have an independent pretheoretical grasp on the notion of metaphysical possibility, one to which an independent space of metaphysically possible worlds may answer.

I think that the first strategy is unpromising, as the logical modality is well-enough behaved. Of course a Quinean might disagree, but I think that such a Quinean then has good reason to be skeptical about the metaphysical modality in addition. The second strategy is also unpromising, as the main data that led philosophers to postulate a distinctive metaphysical modality are the Kripkean data, and these data is accommodated on the picture I have drawn. As for the third strategy, the main further reasons here might be theoretical: for example, the independent reasons to believe in materialism might give one reason to take a distinct metaphysical modality seriously. I think that these reasons count for something, but that they ultimately have to be cashed out by an independent grounding for this modality, in order to have force against the analysis I have suggested.

The fourth strategy is perhaps the most interesting. Here, the idea might be that we have a pretheoretical (and perhaps primitive) grip on the notion of *a way things could have been*, or of other closely related notions in the vicinity. In particular, the idea will be that we have a grip on the notion of the way things *really* could have been, such that it is at least an substantive open question whether all ways that things conceivably could have been are ways that things really could have been.

In response: I think there is good reason to doubt that we have a grip on any such notion, at least insofar as the notion is distinct from rational and nomological notions. We certainly have a grip on various notions of ways things *conceivably* could have been. We also have a grip on the notion of ways things *naturally* could have been, where this notion is tied to how things could have been in our universe with its laws of nature. But there is little reason to think that we have an independent pretheoretical grip on an intermediate notion. Certainly, while there are many uses of modal phenomena in ordinary discourse that invoke broadly epistemic modalities and broadly natural modalities, it is extremely difficult to exhibit uses of modal phenomena in ordinary nonphilosophical discourse that appear to invoke an independent metaphysical modality.

Many philosophers have been persuaded that we possess an independent concept of the metaphysical modality by Kripke's analysis. But Kripke himself simply stipulates that P is (meta-

physically) possible if it might have been the case that P, and goes on to make the case that this notion of possibility comes apart from apriority. That much is plausibly correct, but it does little to support a notion of metaphysical possibility that is independent of conceivability. The Kripkean reasons for believing in a distinctive metaphysical modality are all grounded in the use of subjunctive evaluation of various conceivable situations: about what could and would have been the case, had various conceivable situations obtained. The corresponding notion of metaphysical possibility can be seen as the concept of subjunctive necessity over the space of logically possible worlds. So while Kripke's results make the case that the notion *it might have been that...* (subjunctive) behaves differently from the notion *it might be that* (epistemic), they do little to suggest that the former is independent of conceivability and rational notions. Instead, Kripke's data can all be seen as grounded in the epistemic and subjunctive evaluation of logically possible worlds.

Overall, I think there is good reason to deny that we have an independent concept of metaphysical modality to which a separate space of worlds could answer. The believer in strong necessities must embrace a modal dualism, with distinct and independent metaphysical modalities, and distinct and independent spaces of worlds that answer to them.

The picture I have sketched is a sort of modal monism, with a single primitive space of worlds, along with two sorts of evaluation of sentences over this space. The believer in strong necessities, by contrast, must embrace a modal dualism, with distinct primitive spaces of logically and metaphysical possible worlds. I think that this second primitive is an invention; nothing in our conceptual system requires it. It is a primitive that answers to no-one and does no work.

It is also worth noting that if we postulate a metaphysical modality that is independent of conceivability, the epistemology of modality becomes quite problematic. Certainly, the most widely used route to modal conclusions goes via conceivability. The picture I have given gives an explanation of why this should be so. But if metaphysical modality involves an independent primitive, then it becomes quite unclear why conceivability should be any guide to it at all? Why should not there be just one metaphysically possible world, or 37? An proponent of a distinct modality might postulate principles that entail the existence of a relatively broad space, but now the issue recurs under the guise of the question: why should we accept those principles, if metaphysical modality is primitive? By contrast, the picture on which the metaphysical modality is grounded in the logical modality yields a simple explanation, and a simple epistemology.[4]

[4]Chappell (2006) addresses this challenge by sketching a picture of modality that is available to those who accept a primitive concept of metaphysical modality while simultaneously grounding the use of conceivability. This picture is grounded in two principles: "presumption of possibility", holding that a world-candidate is metaphysically possible

On the picture I have sketched, both the rational modal concepts (rational entailment, apriority, conceivability) and the metaphysical modal concepts (possibility, necessity, property) can be seen as part of a single family. The connection between them is subtle, but both are grounded in the rational domain. The result is modal rationalism in more senses than one: a priori access to modality, and constitutive ties between the modal and rational domains.

Of course there is much more to say about all of this. But one can discern the outlines of an account on which the link between conceivability and possibility is grounded in the rational roots of our modal concepts.

11 Other Anti-Materialist Arguments

12 (1) The Knowledge Argument

The knowledge argument (Jackson 1982), concerning Mary in her black and white room who does not know what it is like to see red, is sometimes put as a simple deductive argument:

(1) Mary knows all the physical facts

(2) Mary does not know all the phenomenal facts

(3) Some phenomenal facts are not physical facts.

However, I think it is obvious that this simple argument on its own cannot defeat physicalism. One can put the point as follows. The term "physical facts" is ambiguous between *narrow physical* facts— facts about some delimited domain such as microphysics (or microphysics, chemistry, and biology, or whatever) and *broadly physical facts*—including high-level physical facts that are not themselves narrowly physical facts, but that supervene metaphysically on those facts. In the Mary scenario as described, it is stipulated that Mary knows all the *narrowly physical* facts. But materialism requires only that phenomenal facts are *broadly physical* facts. So if "physical facts" in the above argument refers to narrowly physical facts, then the conclusion does not entail the falsity of materialism. And if "physical facts" refers to broadly physical facts, then premise (1)

unless there is an explanation of why it is not, and a "consistency principle" holding that any such explanation must be grounded in a priori coherence. These principles are intended to be a priori but substantive, so that they are compatible with the primitivist picture. Of course one can still raise the question of how we can know these principles to be true. But in any case, Chappell's picture offers a useful alternative grounding for modal rationalism and for CP, one that is available to those who accept a distinct and primitive metaphysical modality.

is unsupported by the thought-experiment (since it cannot be taken for granted that Mary knows all the broadly physical facts), and it begs the question against the materialist who holds that phenomenal facts are broadly physical facts.

To avoid this problem, the knowledge argument must be formulated in terms of deducibility and necessitation. One connection between narrowly and broadly physical facts is that the latter are the facts necessitated by the former. This observation on its own does not bridge the gap. But at this point we can observe that the Mary scenario supports something slightly stronger than the claim that Mary knows all the narrowly physical facts. Crucially (as emphasized in Jackson's own exposition), Mary is in a position to know all facts *deducible* from the narrowly physical facts (by a priori reasoning), as she can be taken to be an ideal a priori reasoner. But even then she will not know what it is like to see red. The premises then entitle us to the conclusion that phenomenal facts are not *deducible* from narrowly physical facts. This is now at least in the vicinity of the required conclusion that phenomenal facts are not necessitated by narrowly physical facts (and so are not broadly physical facts). To cross the gap explicitly, we can build in one more premise, yielding the following:

(1) Mary is in a position to know all facts deducible from the narrowly physical facts.

(2) Mary is not in a position to know all the phenomenal facts.

(3) If a phenomenal fact is not deducible from the narrowly physical facts, it is not necessitated by the narrowly physical facts.

(4) Not all phenomenal facts are necessitated by the narrowly physical facts.

I think this is the best way to use the knowledge argument scenario to mount an argument against materialism, while staying close to Jackson's original intentions. One can bring out its relationship to the arguments discussed earlier by making some cosmetic changes to simplify the argument. First, we can talk of truths rather than facts, and represent the narrowly physical truths by P and the relevant phenomenal truth by Q. Second, we can identify deducibility with a priori entailment. Third, we can combine premise (1) and (2) into a single premise. Fourth, we can build in the connection between (4) and materialism explicitly. This yields the following argument:

(1) $P \supset Q$ is not a priori

(2) If $P \supset Q$ is not a priori, $P \supset Q$ is not necessary

(3) If $P \supset Q$ is not necessary, materialism is false.

(4) Materialism is false.

Reformulated this way, the argument has a familiar structure. Its structure is more or less the same as that of the conceivability argument at the beginning of the paper. In fact, the first premise is equivalent to the claim that $P\&\neg Q$ is ideally negatively conceivable, and the second premise is equivalent to the claim that this conceivability entails that $P\&\neg Q$ is possible. So we can see the knowledge argument, formulated this way, as equivalent to a version of the conceivability argument formulated using negative conceivability. Understood this way, the role of the Mary scenario is to give support to the negative conceivability claim in the first premise.

There is one small but important difference between the knowledge argument cast in terms of deducibility and necessitation (I will just call this the knowledge argument, though one could also call it the deducibility argument) and the negative conceivability argument. The premise that $P \supset Q$ is not a priori is slightly stronger than the premise that Mary is not in a position to know Q by a priori reasoning from P. The reason is that the latter premise could be made true by Mary's being unable to acquire the *concepts* involved in Q (such as the concept of phenomenal redness) from inside her black-and-white room. This leaves the way open for the under-discussed *missing-concept* reply to the knowledge argument, which allows that $P \supset Q$ is a priori, but holds that Mary is unable to perform the inference inside the room because she lacks the concept. To defeat this reply, one needs an argument for the stronger claim that $P \supset Q$ is not a priori, for example by arguing that even when Mary has the relevant phenomenal concepts, she cannot always deduce the relevant phenomenal truths (about other people, for example) from Q. But now we are back to a version that is equivalent to the negative conceivability argument.

Comparing the arguments, we can say that the knowledge argument has the weakest first premise, the negative conceivability argument has a somewhat stronger first premise, and the positive conceivability argument has a stronger first premise again. This means that the premise behind the knowledge argument is correspondingly slightly easier to accept than the other two. This difference is reflected in the fact that some philosophers accept the claim about Mary's new knowledge while being uncertain about the conceivability claims. But this advantage is balanced by an inverse relationship in the strength of the premises required to get from the first premise to the failure of materialism. Here, the second premise of the positive conceivability argument is the weakest, that of the negative conceivability argument is somewhat stronger, and the corresponding premise of the knowledge argument (i.e. premise (3) above linking deducibility and necessitation) is stronger again. The missing-concept reply yields an objection to this premise of the knowledge

argument, but not to the corresponding premise of the conceivability argument. And we have seen that the second premise of the positive conceivability argument (an instance of CP+) is weaker than that of the negative conceivability argument (an instance of CP-), in that various potential counterexamples to CP- may yield objections to the latter but not the former.

All three arguments have their place. In my view, the first premises of each are extremely plausible, so it makes sense to focus especially on the argument with the weakest second premise, namely the positive conceivability argument. But for anyone who has doubts about the positive conceivability of zombies or inverts, the negative conceivability argument provides a somewhat weaker first premise that may be easier to accept, along with a second premise that is still plausible. As for the original knowledge argument, I think its highly compelling first premise serves most effectively as part of an argument for the first premise of the negative conceivability argument, although of course this requires some additional work to respond to the missing-concept reply.

In any case, once the knowledge argument is formulated in the version that is equivalent to the negative conceivability argument, the usual dialectic ensues. Most existing replies to the knowledge argument can equally be seen as replies to this argument. For example, the ability reply (holding that knowledge of what it is like to see red involves knowledge-how, not knowledge-that) can be seen as denying the claim that there are any phenomenal truths Q such that $P \supset Q$ is not a priori. And the old-fact reply, holding that Mary learns a physical fact that she already knew under a new phenomenal mode of presentation, can be seen as denying premise (2). This reply is usually inspired by analogies with cases like the water/H_2O case and the Hesperus/Phosphorus case, just as responses to the original premise (2) were.

To respond to old-fact reply and to these analogies, one needs a link between apriority and necessity that is plausible even in light of these cases. Once again, this link can be built through the two-dimensional analysis. In effect, one can replace premise (2) above by the claim that if $P \supset Q$ is not a priori, $P \supset Q$ is not 1-necessary. Then one can add further premises connecting this to the claim that $P \supset Q$ is 1-necessary (with the loophole for Russellian monism, a loophole that is just as wide open in the original knowledge argument) and from there to the conclusion that materialism is false. In my view, the resulting two-dimensional analysis of the knowledge argument is probably the most powerful way to support it.

A related way to develop the knowledge argument in response to the old-fact reply invokes what we might call the New Fact Thesis (Lockwood 1989, pp. 136-37, Chalmers 1996, pp. 141-2, and Thau 2002, p. 127):

> *New Fact Thesis*: Whenever one gains new knowledge of an old fact, one simultaneously gains knowledge of a new fact.

So when Lois, who know that Superman can fly, learns that Clark can fly (old fact, new way) she also learns that someone working for the Daily Planet can fly (new fact). Strictly speaking, to make this thesis connect with the deducibility argument and to accommodate exceptions in the case of indexicals, one should invoke a Modified New Fact Thesis (Chalmers 2005, p. xx):

> *Modified New Fact Thesis*: Whenever one gains new non-indexical knowledge not deducible from previous knowledge, one simultaneously gains new knowledge (or becomes in a position to gain knowledge) of some fact that is not necessitated by the previously known facts.

Then this thesis combines with the other premises of the deducibility argument to yield an argument against materialism.

I think that the result is the most powerful nontechnical version of the knowledge argument. The Modified New Fact Thesis is roughly equivalent in strength to the CP- thesis, and any counterexamples to one are likely to be counterexamples to the other. I think that the best explanation of the truth of the New Fact Thesis is given by the two-dimensional analysis of belief and knowledge. Nevertheless, this version of the argument does not require any special technical apparatus, so it has the advantage of avoiding any objections to the apparatus itself, as well as being somewhat more accessible. So again, these two versions of the argument can work well together.

13 (2) The property dualism argument

The property dualism argument stems from J.J.C. Smart's paper advocating the mind-brain identity theory, and has more recently been developed by Stephen White (1986; 2006). Smart attributes to Max Black the objection that if mental states and physical states are contingently identical, they must still be picked out via different properties. If so, then mental terms and physical terms will be associated with distinct properties. In more contemporary terms, the underlying thesis might be put as follows:

> *Distinct-Property Thesis*: When an identity claim '$a = b$' is not a priori, 'a' and 'b' pick out their referents via distinct properties.

One might then formalize the argument as follows, where 'q' is an arbitrary mental term:

(1) For all physical terms 'p', '$p = q$' is not a priori

(2) If '$p = q$' is not a priori, 'p' and 'q' pick out their referent via distinct properties.

(3) For all physical terms 'p', 'q' picks out its referent by a property distinct from the property that picks out the referent of 'p'.

Here the conclusion is not yet the denial of materialism. One could get closer by augmenting the argument with the premise that physical terms have the same property as referent and as reference-fixer. The argument will then warrant the conclusion that the reference-fixer for 'q' is distinct from any physical property. Of course an opponent may question the extra premise for reasons familiar from the discussion of physical terms in section 3, but as before, a proponent might respond by arguing that this loophole leads only to a Russellian version of the identity theory. Likewise, an opponent may suggest that the argument only warrants the claim that the reference-fixer for 'q' is distinct from any narrowly physical property, not from any broadly physical property. A proponent might reply by invoking the thesis that any broadly physical property can be expressed by a physical term.

The argument can be developed in various ways; see White (1986; 2006) and Block (2006) for versions somewhat different from the above. But it is clear that what is crucial to each is the Distinct-Property Thesis embodied in premise 2. So here I will concentrate on that thesis, and on its relationship to the two-dimensional analysis.

The Distinct-Property Thesis is closely related to the two-dimensional principle CP-. The thesis entails that when 'a=b' is not a priori, so that 'a!=b' is negatively conceivable, 'a!=b' will be 1-possible, which entails that 'a' and 'b' have distinct primary intensions. Primary intensions are not properties, but there is a close relationship. Any property corresponds to an intension (over uncentered possible worlds), picking out those individuals that have the property in a given world. Some theorists identify the property with the associated intension, and most agree that a property at least determines in an intension. If so, then where there are distinct intensions, there are distinct properties.

The two-dimensional analysis also provides a natural way to determine the property, or at least the intension, associated with a given expression. The intension can be understood as a function that picks out what an expression picks out at a world, under the hypothesis that the world in question is actual. This understanding fits naturally with the understanding in White (2006), where the distinct-property thesis is motivated by the claims that (i) when an identity is a posteriori, it can rationally be disbelieved, and (ii) for a rational belief, there will always be a possible world

that rationalizes the belief: roughly, a world such that if that world were actual, the belief would be true. If we associate expressions with intensions as above, then these two claims (along with plausible auxiliary premises) entail that when 'a=b' is not a priori, 'a' and 'b' will be associated with distinct intensions.

One important difference is that where properties correspond to intensions over uncentered worlds, primary intensions are intensions over centered worlds. This difference reflects a defect in the distinct-property thesis as it stands. Indexical identity claims, such as 'I am David Chalmers', are not a priori, but it is hard to find a reference-fixing property associated with indexicals such as 'I' and 'here', and there is no obvious uncentered world that rationalizes the denial of such a claim. To fix this problem, one needs to either exclude indexicals from the scope of the distinct-property thesis, or modify the thesis so that it invokes primary intensions (or perhaps reference-fixing relations) rather than properties, with centered worlds playing the role of possible worlds.

Modified in this way, the distinct-property thesis is equivalent to thesis CP-. The ensuing dialectic is also quite similar. For example, a type-B materialist might respond to White's argument by saying that rational beliefs require an *epistemically* possible situation in which the belief is true, but do not require a *metaphysically* possible situation. In the terms used earlier, this comes to the claim that the identity claim in question may be false relative to some *scenario* (considered as actual), but that this scenario does not correspond to a metaphysically possible world. In effect, this response requires the existence of strong necessities, with the ensuing dialectic as before.

Likewise, Block (2006) responds to the property dualism argument by suggesting that a posteriori identity claims always involve distinct *cognitive* modes of presentation (CMoPs), but need not involve distinct *metaphysical* modes of presentation (MMoPs, or properties). In terms of the two-dimensional framework, CMoPs might be seen as corresponding to epistemic intensions defined over scenarios, while MMoPs might be seen as primary intensions over centered worlds. The residual question is then, once again, the question of whether there is a centered world for every scenario. If there is, then distinct CMoPs will always entail distinct MMoPs. In any case, primary intensions are well-suited to play the role of MMoPs, and the residual issue once more comes down to the question of whether there are strong necessities.

14 (3) Descartes' argument from disembodiment

One version of Descartes' conceivability argument runs as follows. Here B can stand for my body, or for any physical thing.

(1) It is conceivable that I am not B.

(2) If it is conceivable that I am not B, it is possible that I am not B.

(3) If it is possible that I am not B, then I am not B.

(4) I am not B.

Here, the first premise is based on an intuition about disembodiment: it seems conceivable that I could exist without my body existing. The second premise is an instance of a general thesis connecting conceivability and possibility, and the third premise reflects the claim that any physical thing is essentially that physical thing. The conclusion, generalized, is that I am not any physical thing, but instead am a nonphysical thing.

The soundness of this argument is often doubted, and the reasons for this doubt can be expressed straightforwardly in the current framework. The sense of conceivability in which premise (1) is plausible is primary (positive or negative) conceivability. To connect with this interpretation of premise (1), premise (2) must be interpreted as involving 1-possibility. For premise (3) to be plausible, on the other hand, it must be interpreted as involving 2-possibility: when identity statements involving rigid designators are true (or false), they are 2-necessary (or 2-impossible). But if premise (2) involves 1-possibility and premise (3) involves 2-possibility, then the argument is not valid.

One might try to find a univocal reading of the argument, but such a reading will not succeed. If premise (1) is invoking secondary conceivability, then it may fail: secondary conceivability depends on how things turn out empirically, and if it turns out that I am physical, my disembodiment will not be secondarily conceivable. If premise (1) invokes primary conceivability but premise (2) invokes 2-possibility, then it will fail, as primary conceivability does not entail 2-possibility. If premise (3) invokes 1-possibility, then it will fail, as true identity statements need not be 1-necessary. So we arrive at a familiar diagnosis of Descartes' arguments: the sort of conceivability that he is entitled to does not ground the sort of possibility that he needs.

In this case, it appears that there is no straightforward way to fix up the argument in the way that we fixed up the conceivability argument earlier. It may be that there is some way to repair the argument, but if so, it will require more than the two-dimensional tools here.

15 (4) Kripke's modal argument

The anti-materialist argument that is most closely related to the two-dimensional argument is Kripke's modal argument against the identity theory. Kripke's argument can put as follows. Let 'p' stand for pain and 'c' be a term for C-fiber firing. Then

(1) '$p = c$' is apparently contingent.

(2) If '$p = c$' is apparently contingent, then there is a world with a being in an epistemic situation that is qualitatively identical to mine in which a corresponding statement is false.

(3) If there is a world with a being in an epistemic situation that is qualitatively identical to mine in which a statement corresponding to '$p = c$' is false, then there is a world at which '$p = c$' is false.

(4) If there is a world at which '$p = c$' is false, then '$p = c$' is false.

(5) '$p = c$' is false.

Here, (2) expresses the model that Kripke thinks works in all the usual cases of "apparently contingent" identity statements. For example, 'heat=molecular motion' is apparently contingent (although true and necessary), and its apparent contingency is explained by a world with a being in my epistemic situation for whom a corresponding statement would be false. In particular, in a world where heat sensations are not produced by molecular motion, such a being's utterance of 'heat=molecular motion' would be false. This model turns on their being a difference between being felt as heat and being heat. But as (3) suggests, Kripke holds that this model does not apply in the case of pain. Anything that is felt as pain is pain, and any epistemic situation in which something is felt as pain is an epistemic situation in which there is pain. So if a counterpart statement is false at a world, because C-fibers are not felt at pain in that world, then the original statement will also be false at that world, because C-fibers cannot be pain at that world. So C-fibers cannot be pain anywhere.

The first premise can be seen as saying that it is conceivable in one sense that pain is not C-fibers. Kripke does not say a great deal about the apparent contingency of a statement, but he does say that it goes along with the sense that the statement could have turned out to be false. It seem reasonable to identify this with something like primary positive conceivability of that statement's negation. Furthermore, the second premise is closely related to the claim that 'p=c' is 1-possible.

Let us say that the *Kripke intension* of an utterance is defined at all centered worlds where the being at the center is in the same epistemic situation and making a corresponding utterance, and is true at such a world if that utterance is true. And let us say that a statement is *K-possible* if its Kripke intension is true at some centered world. Then premise (2) says that if a true identity statement is apparently contingent, it is K-possible. Putting all this together, the argument can be seen as follows:

(1) '$\neg(p = c)$' is 1-conceivable

(2) If '$\neg(p = c)$' is 1-conceivable, '$p = c$' is K-possible.

(3) If '$\neg(p = c)$' is K-possible, '$\neg(p = c)$' is possible.

(4) If '$\neg(p = c)$' is possible, '$p = c$' is false.

(5) '$p = c$' is false.

Interpreted this way, Kripke's argument is quite close in structure to the two-dimensional argument I have given. One obvious difference is that Kripke's argument only applies to the type identity theory, not to materialism in general. But it is clear that one might extend it, for example by replacing the identity claim '$p = c$' with a conditional such as $P \supset Q$ and making some adjustments. Another difference is that Kripke's argument does not take into account the loophole that allows in Russellian monism, which (as Maxwell 1978 observed) can itself be seen as a sort of identity theory on which 'p=c' is true, because both pain and C-fiber firing are identical to a hidden intrinsic property that underlies the dispositions associated with C-fiber firing. This is because (as Maxwell also observes) Kripke's argument for (3) takes into account the possibility of an appearance-reality distinction for 'pain', but overlooks the possibility of a corresponding distinction for 'C-fiber firing'. When one takes this possibility into account room for a Russellian identity theory is left open. But as before, one can always accommodate this loophole by building in Russellian monism as a disjunct, in the consequent of premise (3) and in the conclusion.

I think that the biggest problem with Kripke's argument lies with the second premise. It is not true in general that if the negation of an identity statement is 1-conceivable, it is K-possible. For example, I might introduce 'Bill' as a descriptive name that rigidly designates whatever phenomenal quality I am currently experiencing at the center of my visual field. As it happens, Bill is phenomenal blueness. But this statement seems apparently contingent in the way that any ordinary identity statement does. Like other such statements, 'Bill is not phenomenal blueness' is 1-conceivable ('Bill is the quality in the center of my visual field' may be a priori, but 'Bill is

phenomenal blueness' is not). But 'Bill is not phenomenal blueness' is not K-possible. Any being in a qualitatively identical epistemic situation to mine will be experiencing phenomenal blueness in the center of their visual field. For any such being, a statement corresponding to 'Bill is not phenomenal blueness' will be false. So 'Bill is phenomenal blueness' is K-necessary, and the generalization of premise (2) is not true in general: 1-conceivability does not entail K-possibility.

This is a serious problem for Kripke's formulation of the argument. It is especially problematic because Kripke's reasoning in the case of pain and C-fibers seems to rely on just the same features that are present in the case of phenomenal blueness and Bill. It is reasonable for an opponent to say that where this argument is concerned, the identity between pain and C-fiber firing is on a par with the identity between phenomenal blueness and Bill.

One can get around this problem, but it requires dropping Kripke's framework of corresponding epistemic situations and corresponding statements. To eliminate the problem, we need to move from Kripkean intensions to primary intensions. Kripkean intensions resemble primary intensions in some respects, but they are defined quite differently: Kripkean intensions are defined in terms of the truth of certain counterfactual statements at counterfactual worlds, whereas primary intensions are defined in terms of the epistemic status of certain conditionals at our world. (In the terminology of Chalmers 2004a, Kripkean intensions are defined *contextually*, while primary intensions are defined *epistemically*.) Correspondingly, Kripkean intensions behave like primary intensions in some situations, but not all situations. The case of 'Bill=phenomenal blueness' is one of these. The Kripkean intension of my utterance of this statement is true at all centered worlds where it is defined, but the primary intension is not. In particular, the primary intension of this statement will be false at a centered world in which the individual at the center is experiencing phenomenal redness. Given the way 'Bill' functions, if I accept that my centered world is such a world, then I should accept that Bill is phenomenal redness. So while 'Bill is not phenomenal blueness' is not K-possible, it is 1-possible. The case falsifies an entailment from the relevant sort of conceivability to K-possibility, but it is quite compatible with an entailment from conceivability to 1-possibility.

I conclude that Kripke's argument can be modified to yield a general argument against materialism from plausible premises, and that the best way to do so is to invoke the two-dimensional framework in the way that I have discussed. This is not too surprising, as the argument can itself be considered a sort of refinement of Kripke's original argument. Doing things two-dimensionally also has the advantage that many objections to Kripke's original argument can be seen straightforwardly to fail, though of course there are new objections in their place. In any case, the two-dimensional connection between the epistemic and the ontological domain seems once again to be

the central locus on which many of these issues turn.[5]

(5) *The semantic stability argument*. Bealer (1994; 2002) gives a closely related modal argument. The key premise of Bealer's argument is something along the lines of the following:

> (SS) If S is semantically stable, then if one has an a priori stable intuition that S is possible, then S is possible.

Here a term is semantically stable iff, necessarily, in any language group in an epistemic situation qualitatively identical to ours, the expression would mean the same thing. So, for example, 'two' is semantically stable but 'water' is not. The restriction to semantic stability is intended to rule out Kripke-style a posteriori necessities such as 'water is $\mathrm{H_2O}$'. As for the rest of (SS), Bealer's "intuition that S is possible" seems to come to roughly the same thing as positive conceivability, and his "a priori stable" intuition of possibility comes to something like ideal positive conceivability.

An immediate problem is that Bealer's key premise (SS) is vulnerable to the same counterexamples as the Kripkean premise above. For example, the term 'Bill' discussed above is semantically stable according to Bealer's criterion, but 'Bill=phenomenal blueness' is an a posteriori necessity. Correspondingly, there is a modal intuition that 'Bill is not phenomenal blueness' is possible, but it is not possible. Something similar applies to 'L', a rigid designator for the number of languages actually spoken. 'L' is semantically stable by the definition above, but '$L > 0$' is an a posteriori necessity.

To get around this problem, one can replace Bealer's notion of semantic stability by the related notion of semantic neutrality. A term T is semantically neutral iff its modal profiles are determined a priori. In two-dimensional terms, the secondary intension of T must be determined a priori: that is, for all scenarios $v1$ and $v2$ and all worlds w, the extension of T at $(v1, w)$ and at $(v2, w)$ coincide. Then the reformulated thesis, in our language, says:

> (SN) If S is semantically neutral, then if S is conceivable, S is possible.

Here 'conceivable' stands for ideal positive conceivability, and 'possible' can stand for ordinary metaphysical possibility, as primary/secondary distinctions do not come into play when S is semantically neutral. The counterexampes to SS involving 'Bill' and 'L' are not counterexamples to SN, as 'Bill' and 'L' are not semantically neutral: their modal profile is determined a posteriori.

SN is an immediate consequence of thesis CP. CP tells us that when S is conceivable it is 1-possible, and when S is semantically neutral, S is 1-possible iff it is 2-possible. On the face of it,

SN is somewhat weaker than CP, as it makes no commitment about the modal status of non-neutral sentences. Correspondingly, SN is perhaps slightly easier to accept, both because it is weaker and because its formulation requires less technical apparatus. At the same time, CP is more powerful in allows an analysis of issues involving non-neutral expressions. In any case, the theses are clearly closely related, and the most important potential counterexamples to CP discussed earlier are also potential counterexamples to SN.

Because most physical expressions are arguably not semantically neutral, one cannot apply SN directly to them, and in particular one cannot use it directly to draw the conclusion that $P\&\neg Q$ is possible. But one can argue against many identity claims indirectly. For example, Bealer argues against 'pain is C-fibre firing' by applying SS to the conceivability of 'A being is in pain and has no part with more than n parts', for large n such that a C-fibre necessarily has more than n parts. Bealer himself only uses the strategy to argue against identity physicalism, appealing to different considerations to argue against other forms. But one could go a step further to obtain a more general argument. One might apply SN to the conceivability of $S\&\neg Q$, where S is a semantically neutral characterization of the structure of microphysics, yielding (when combined with the premise that Q is semantically neutral) the conclusion that $S\&\neg Q$ is possible. One can then argue that if $S\&\neg Q$ is possible, then either $P\&\neg Q$ is possible or Russellian monism is true. What results will be a dialectic quite similar to the dialectic discussed above.[6]

[6]A recent anti-materialist argument by Nida-Rümelin (2006) raises related issues. Nida-Rümelin's key premise is a principle of cognitive transparency (CT), holding that anyone who grasps a property P via two distinct concepts can in principle find out without further empirical investigation that those concepts are necessarily coextensive. Here one grasps a concept when one is in a position to know its modal profile, given one's background knowledge (so one grasps the concept *water* iff one knows that water is essentially H_2O). Nida-Rumelim uses the two-dimensional apparatus to define and analyze the notion of grasping, but at the same time she suggests that (CT) is weaker and less controversial than (CP).

(CT) is closely related to (SN), however. The connection derives from the fact that according to Nida-Rümelin's formal definitions, anyone who possess a semantically neutral concept grasps the concept. Applying (CT) to this special case and making natural assumptions, one is led to (CT'): Any true property identity 'p=q' involving semantically neutral expressions is knowable a priori. (CT') immediately yields the following special case of (SN): When (semantically) neutral '$\neg(p = q)$' is ideally negatively conceivable, it is true, and therefore possible. Correspondingly, potential counterexamples to (SN) will typically yield counterexamples to (CT), as will the major potential counterexamples to (CP). For example, someone such as Loar (1990/1997), who holds that there are true a posteriori property identities involving semantically neutral physical and phenomenal concepts, will deny (CT') and (CT). ((CT') is near-equivalent to the "semantic premise" that Loar rejects, which is in turn closely connected to the Distinct-Property Thesis discussed earlier; see Chalmers 2004 for more discussion here.) Still, the plausibility of (CT) might be seen as providing one more reason to reject these views. For more on this matter, see consc.net/mnr.html.

References

Alter, T. forthcoming. On the conditional analysis of phenomenal concepts. Philosophical Studies.

Ashwell, L. 2003. *Conceivability and Modal Error*. MA Thesis, University of Auckland.

Austin, D.F. 1990. *What's the Meaning of "This"?* Ithaca, NY: Cornell University Press.

Balog, K. 1999. Conceivability, possibility, and the mind–body problem. Philosophical Review 108:497-528.

Bealer, G. 1994. Mental properties. *Journal of Philosophy* 91:185-208.

Bealer, G. 2002. Modal epistemology and the rationalist renaissance. In (T. Gendler & J. Hawthorne, eds) Conceivability and Possibility. Oxford University Press.

Block, N. 2006. Max Black's objection to mind-brain identity. In (T. Alter & S. Walter, eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.

Block, N. & Stalnaker, R. 1999. Conceptual analysis, dualism, and the explanatory gap. Philosophical Review.

Chalmers, D.J. 1996. The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Chalmers, D.J. 1999. Materialism and the metaphysics of modality. Philosophy and Phenomenological Research.

Chalmers, D.J. & Jackson, F. 2001. Conceptual analysis and reductive explanation. Philosophical Review.

Chalmers, D.J. 2002. Does conceivability entail possibility? In (T. Gendler & J. Hawthorne, eds) Conceivability and Possibility. Oxford University Press.

Chalmers, D.J. 2003a. Consciousness and its place in nature. In (S. Stich & F. Warfield, eds) *The Blackwell Guide to the Philosophy of Mind*. Blackwell. [consc.net/papers/nature.html]

Chalmers, D.J. 2003b. The content and epistemology of phenomenal belief. In (Q. Smith & A. Jokic, eds) *Consciousness: New Philosophical Essays*. Oxford University Press.

Chalmers, D.J. 2004a. Epistemic two-dimensional semantics. Philosophical Studies 118:153-226.

Chalmers, D.J. 2004b. Imagination, indexicality, and intensions. Philosophy and Phenomenological Research 68:182-90.

Chalmers, D.J. 2005. Phenomenal concepts and the knowledge argument. In (P. Ludlow, Y. Nagasawa, & D. Stoljar, eds) *There's Something About Mary*. MIT Press.

Chalmers, D.J. 2006a. Phenomenal concepts and the explanatory gap. In (T. Alter & S. Walter, eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.

Chalmers, D.J. 2006b. Two-dimensional semantics. In (E. Lepore & B. Smith, eds) *The Oxford Handbook to the Philosophy of Language*. Oxford University Press.

Chalmers, D.J. forthcoming. The nature of epistemic space.

Fine, K. 2002. The varieties of necessity. In (T. Gendler & J. Hawthorne, eds) *Conceivability and Possibility*. Oxford University Press.

Hawthorne, J. 2002. Advice to physicalists. *Philosophical Studies* 109:17-52.

Hawthorne, J. 2004. The deeply contingent a priori. *Philosophy and Phenomenological Research*.

Hawthorne, J. 2006. Dancing qualia and direct reference. In (T. Alter & S. Walter, eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.

Hill, C.S. 1997. Imaginability, conceivability, possibility, and the mind–body problem. Philosophical Studies 87:61-85.

Hill, C. S. & McLaughlin, B. P. 1998. There are fewer things in reality than are dreamt of in Chalmers' philosophy. Philosophy and Phenomenological Research.

Ismael, J. 1999. Science and the phenomenal. *Philosophy of Science*.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136.

Jackson, F. 1998. *From Metaphysics to Ethics*. Oxford University Press.

Kirk, R. 1999. Why there couldn't be zombies. Aristotelian Society Supplement 73:1-16.

Kripke, S.A. 1980. *Naming and Necessity*. Harvard University Press.

Lewis, D. 1986. *On the Plurality of Worlds*. Blackwell.

Loar, B. 1997. Phenomenal states (second version). In (N. Block, O. Flanagan, & G. Güzeldere, eds) *The Nature of Consciousness*. MIT Press.

Loar, B. 1999. David Chalmers' *The Conscious Mind*. Philosophy and Phenomenological Research 59:464-71.

Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Blackwell.

Lynch, M. forthcoming. Zombies and the case of the phenomenal pickpocket. *Synthese*.

Marcus, E. 2004. Why zombies are inconceivable. Australasian Journal of Philosophy 82:477-90.

Marton, P. 1998. Zombies vs. materialists: The battle for conceivability. *Southwest Philosophy Review* 14:131-38.

Maxwell, G. 1979. Rigid designators and mind-brain identity. Minnesota Studies in the Philosophy of Science 9.

Nida-Rümelin, M. 2006. Grasping phenomenal properties. In (T. Alter & S. Walter, eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism. Oxford University Press*.

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

Schiffer, S. 2003. Two-dimensional semantics and propositional attitude content. In *The Things We Mean*. Oxford University Press.

Shoemaker, S. 1999. On David Chalmers' *The Conscious Mind. Philosophy and Phenomenological Research* 59:539-44.

Sidelle, A. 2002. On the metaphysical contingency of laws of nature. In (T. Gendler & J. Hawthorne, eds) *Conceivability and Possibility*. Oxford University Press.

Soames, S. 2004. *Reference and Description: The Case Against Two-Dimensionalism*. Princeton University Press.

Stoljar, D. 2001. The conceivability argument and two conceptions of the physical. Philosophical Perspectives 15:393-413.

Sturgeon, S. 2000. *Matters of Mind: Consciousness, Reason, and Nature*. Routledge.

Thau, M. 2002. *Consciousness and Cognition*. Oxford University Press.

Thomas, N. J. T. 1998. Zombie killer. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) Toward a Science of Consciousness II. MIT Press.

van Gulick, R. 1999. Conceiving beyond our means: The limits of thought experiments. In (S. Hameroff, A. Kaszniak, & D. Chalmers, eds) Toward a Science of Consciousness III. MIT Press.

White, S. 1986. Curse of the qualia. Synthese 68:333-68.

White, S. forthcoming. Why the property dualism argument won't go away. Journal of Philosophy.

Worley, S. 2003. Conceivability, possibility and physicalism. Analysis 63:15-23.

Yablo, S. 1998. Concepts and consciousness. Philosophy and Phenomenological Research.

Yablo, S. 1998. Textbook Kripkeanism and the open texture of language. Philosophical Quarterly.

Yablo, S. 2002. Coulda, woulda, shoulda. In (T. Gendler & J. Hawthorne, eds) Conceivability and Possibility. Oxford University Press.