Kripke's Proof That We Are All Intuitive Dualists

David Papineau

1 Introduction

The aim of this paper is to show that we are all in the grip of an intuition of dualism. I shall use Kripke's argument from Lecture III of Naming and Necessity to establish this. I do not think that Kripke's argument proves that dualism is true. But I do think that it demonstrates that dualism has us all in its intuitive grip.

In my view the force of Kripke's argument is little appreciated because it is widely conflated with a significantly different argument, the 'two-dimensional argument against physicalism'. Kripke's argument is much better than this two-dimensional argument. It is easy for contemporary physicalists to answer the two-dimensional argument. But Kripke's original argument calls for a much more complicated response.

Before I get down to details, I would like to explain why this issue matters. Physicalism is commonly held to leave us with an 'explanatory gap' (Levine 1983). Compare putative mind-brain identity claims like pain = C-fibres firing with scientific identity claims like water = H_2O or heat = molecular motion. When we are introduced to claims of the latter sort, and shown the relevant evidence, we have no difficulty in accepting their truth. But things seem different with claims like pain = C-fibres firing. Even if we became persuaded that pains and C-fibre firings always accompany each other, and recognized that this kind of evidence would normally suffice to establish identity, we would still seem to face an explanatory challenge. Why should C-fibre firings yield pain? How can the brain state suffice for the feeling? As Thomas Huxley put the worry over a hundred years ago: 'How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of the Dijn, where Aladdin rubbed his lamp' (1866).

It is widely supposed that this impression of an explanatory gap arises because our pre-theoretical concepts of pain and other conscious states do not allow a priori derivations of mind-brain identities from the physical facts, in the way that concepts like water and heat arguably allow the corresponding derivations of the scientific identities. The implication is that there is something wrong with current physicalism. In order to be successful, physicalism needs to do something more. It needs to come up with some alternative way of conceiving conscious states, some way that will allow us to bridge the explanatory gap.

I have a quite different diagnosis. I think that the so-called 'explanatory gap' is simply a manifestation of an intuitive conviction that dualism is true. It's not that mind-brain identities are hard to explain—they are simply hard to believe. When we consider a putative identity like pain = C-fibres firing, our intuitive reaction is simply that this claim must be false, because pain and C-fibre firings are distinct states. As long as we remain in this state of mind, then of course we will think that there are explanatory questions that have not been answered. Why do C-fibre firings give rise to the extra feeling of pain? What rules out the possibility of their yielding some different feeling, or no feeling at all? These seem obvious questions, yet questions to

which current physicalism offers no good answer. Still, if my diagnosis is right, these are not questions that remain to be answered even after we accept physicalism. Rather, they are a manifestation of the difficulty of accepting physicalism in the first place.

According to my diagnosis, then, the intuitive feeling of an 'explanatory gap' is nothing to do with to the impossibility of deriving identities like pain = C-fibres firing a priori from the physical facts. There are many other examples of true identities which aren't so a priori derivable, yet which we happily accept without any accompanying feeling of a gap. (Block and Stalnaker, 2000, Papineau, 2002, ch 5.) What distinguishes mind-brain identities, and generates the feeling of an intuitive gap, is simply our strong intuition that these identities cannot be true to start with.

The implication is that physicalists are barking up the wrong tree if they think that the apparent 'gap' calls for some new improved version of physicalism, offering extra resources to help us construct a priori derivations of mind-brain identities. That's not why we feel dissatisfied. There is nothing lacking in current physicalism itself. The problem is simply that it is hard to believe. If only we could get ourselves into a frame of mind where we fully believed physicalism, then the appearance of some distinctive gap would dissolve, and we wouldn't feel that something remained to be done.

Have invoked the 'explanatory gap' to explain why the intuition of distinctness matters, I will say nothing more about it in the body of this paper. Henceforth my focus will be on the intuition of distinctness itself. No doubt there is more to say in defence of my view that the appearance of an explanatory gap is due solely to a persistent intuition of distinctness. I take this to be the obvious explanation for the apparent gap, once it is granted that such an intuition of distinctness exists.[1[1]] But a full defence of this claim would need to compare my diagnosis with alternative accounts of the appearance of a gap. Still, as I said, this is not my present concern. It will be enough for this paper if I can persuade readers that the dualist intuition exists, even if they continue to think that there is something more to the 'explanatory gap'.

I have previously argued, in a number of places, that a persistent intuition of distinctness is the underlying source of our puzzlement about the mind-body problem (Papineau, 1993, 1997, 2002), but hitherto I have been more concerned with possible explanations for this intuition, rather than with the existence of the intuition itself. This paper is a change of tack. I shall say nothing here about why we might be prey to a persistent intuition of distinctness. My only concern will be to show that we are. A full understanding of the intuition would obviously involve an explanation of its origin and persistence, but this will not be something that I offer here. (For further discussion of possible explanations, see Melnyk 2003, Papineau 2006.)

Let me mention something else I won't be arguing for. I have said that it is important to recognize the intuition of distinctness because this can help forestall the charge that current physicalism is explanatorily deficient. As this indicates, I take it that there are

---

[1[1]] In support of this claim, consider how the gap is normally introduced—'why do brains "give rise to"/"generate"/"cause"/etc conscious feelings?' These phrases all presuppose that the conscious mind is ontologically distinct from the brain.

good reasons to embrace current physicalism. But this is not something I shall defend here. The positive case for physicalism is familiar, and there will be no need to go into details in this paper. (Cf. Levine 2001 ch 1, Papineau 2002 ch 2.)

A couple of final preliminaries. I have presented physicalism as consisting of identity claims like pain = C-fibres firing. Others prefer to think of physicalism in terms of metaphysically necessary supervenience, rather than identity. This difference is not important for current purposes. All the arguments which follow could as easily (if not as graphically) be phrased in terms of supervenience as identity.

If we do think of physicalism in terms of identity, there is a question about what kinds of physical states will appear on the right hand side of identity claims. Will they be strictly physical states, or physically realized functional states, or what? My illustrative use of the claim that pain = C-fibres firing should not be thought to commit me to any particular view on this issue. I am happy to leave it open exactly which kinds of physical states will feature in the correct formulation of physicalism. Nearly all the arguments which follow are insensitive to this choice (I will mention the issue explicitly when it matters). For the most part, then, the phrase 'C-fibres firings' should simply be understood as a place-holder for the name of whatever physical state turns out to the best candidate to reduce pain, and 'physical' should be understood in a generous sense, so as to include states which metaphysically supervene on physical states as well as strictly physical states.

The rest of this paper will proceed as follows. In the next section I shall rehearse Kripke's original anti-physicalist argument. The following section will distinguish this argument from a more recent argument with which it is often confused, the two-dimensional argument against physicalism. I shall then show, in sections 4 and 5, that the standard physicalist response to the two-dimensional argument is no response to Kripke. The following section will then consider how physicalists can respond to Kripke's actual argument, and will conclude that they have no alternative but to hold that at some level they do not fully believe physicalism. A final section addresses the worry that this physicalist response to Kripke is ad hoc.

2 Kripke's Argument

As I said, I want to use Kripke's anti-physicalist argument from the end of Naming and Necessity to show that we are all in the grip of a dualist intuition. Let me begin by reminding you how this argument goes.

After some preliminaries, Kripke turns to type-type identities like pain = C-fibres firing (p 148). If such an identity obtains, then it obtains necessarily. Even so, claims like pain = C-fibres firing certainly seem contingent. There certainly seem to be metaphysically possible worlds in which C-fibres fire, yet there are no pains. ('Zombie worlds' as we would call them now, though this is not Kripke's terminology.)

Of course, Kripke immediately points out, no clear-headed physicalist will allow that such a situation is really possible. If it is possible to find C-fibre firings without pains, then pains cannot be C-fibre firings, since identity is a necessary relation. All the same, physicalists will be hard put to deny that the divergence of pains from C-fibres

seems possible. Physicalists surely have to admit there is an appearance of contingency, even if they then add that this appearance is misleading, and the divergence is not really possible.

Given this, Kripke then challenges the physicalists to explain why this appearance should arise: given they hold that pains are C-fibre firings, then why does it so much as appear possible that these states should come apart? Kripke holds that physicalists can give no good answer to this question (p 150).

In support of this claim, Kripke contrasts mind-brain identities with scientific identities like heat = molecular motion (pp 151-2). In the latter case too there is an appearance of contingency. It seems possible that molecular motion might not have been heat. But in this case an explanation for the appearance of contingency is readily available, consistently with the truth of the identity. When we suppose that molecular motion might not have been heat, we suppose that molecular motion might not have given rise to sensations of warmth in sentient beings. That is, we suppose that the kind in question might not have given rise to the appearance by which we ordinarily recognize it. This possibility, that the appearance comes apart from the kind, is different from the supposition that the kind itself (heat) comes apart from its scientific nature (molecular motion), and so is perfectly consistent with the original claim that the kinds are strictly identical with those natures.[2[2]]

But this kind of explanation of how a necessary identity can seem contingent will not work in the mind-brain case. Think how the corresponding story would run: in supposing that C-fibre firings might not have been pains, we are supposing that the relevant kind—C-fibre firings, that is, pains—might not have given rise to the sensations—the nasty, hurty feelings—by which we recognize them. But this will not do. Pains can't be pulled apart from their appearances, in the way that water and heat can. A world in which C-fibre firings don't yield hurty feelings isn't a world in which pains lack their normal appearance—it's a world in which pains don't exist at all. So this world is no good for explaining the appearance of contingency. It's not a world in which we have C-fibre firings, that is, pains, but not their appearance. It's simply a world in which we have C-fibres without pains. And the possibility of this world isn't consistent with the supposition that C-fibres are pains. If C-fibres are pains, then there is no room for the possibility that they might not have been themselves (pp 151-3).

I take Kripke to pose a real difficulty for physicalists. In the end, I think that he is quite right to hold that they have no good way to account for the apparent contingency of claims like pain = C-fibres firing, consistently with a whole-hearted commitment to such identities. Whether this amounts to a refutation of physicalism, however, is a further question. Kripke presents his argument as targeted on this conclusion. But it

---

[2[2]] According to Kripke, heat is a rigid designator, and so the claim molecular motion might not have been heat is not itself strictly true; the appearance of contingency arises only because the related claim molecular motion might not causes heat sensations is true. Others hold that heat can be understood so as to make molecular motion might not have been heat itself true. I can skirt round this issue, since neither way of explaining the appearance of contingency is available in the mind-brain case. The crucial issue is not whether heat is a rigid designator, but that it, unlike concepts like pain, has its reference fixed as the kind that contingently displays a certain appearance.

seems to me that there is room for physicalists to respond to his analysis by admitting that they are less than fully committed to mind-brain identities at an intuitive psychological level, while continuing to insist on such identities at a theoretical level.

Still, we can leave the precise consequences of Kripke's argument until later. First we need to show that it does indeed pose a real problem for physicalists. Many contemporary philosophers will be surprised that I am prepared to concede this much cogency to Kripke's argument. This is because it is widely supposed that a crucial premise in Kripke's argument will simply be denied by contemporary physicalists.

3 The Two-Dimensional Argument

I think that this reaction rests on a confusion between Kripke's argument and a different anti-physicalist argument. Disentangling the two arguments is crucial to appreciating the force of Kripke's argument. Over the last decade David Chalmers and Frank Jackson have developed what I shall call 'the two-dimensionalist argument' against physicalism (Jackson, 1993, 1998, Chalmers 1996). This argument is widely supposed to be effectively the same argument as Kripke's. However, as I shall show, the two arguments are importantly different. In particular, where the two-dimensional argument does indeed rest on an assumption that contemporary physicalists will simply deny, Kripke's own argument does not require this assumption. In consequence, the standard physicalist response to the two-dimensional argument leaves Kripke's argument untouched.

The two-dimensional argument against physicalism can be developed in a number of different ways (cf. Chalmers 2006). Fortunately the differences do not matter for the points I wish to make. It will be enough for present purposes to explain the argument at a fairly intuitive level.

According to contemporary physicalism, phenomenal mind-brain identities like pain = C-fibres firing are a posteriori necessities.[3[3]] The two-dimensional argument objects that phenomenal mind-brain identity claims cannot be a posteriori necessities, because a posteriori necessity is characteristically due to 'semantic instability', but phenomenal concepts are not semantically unstable.

Let me unpack this. The central idea here is that of semantic instability[4[4]]. Intuitively, a concept is semantically unstable if it refers to different things depending on how the actual facts turn out. Thus, if the reference of water is fixed as that stuff, whatever it is, that is odourless, colourless and falls from the skies, then water is semantically unstable, in that this concept will refer to different things, depending on

---

[3[3]] Here and throughout I shall assume that contemporary physicalists adopt the 'type-B' view that the interesting mind-brain identities involve special phenomenal concepts that are a priori distinct from any physical or functional concepts. Logical behaviourists and analytical functionalists adopt the alternative 'type-A' view that our concepts of conscious states do render the relevant mind-brain identities a priori. I take type-A physicalism to be untenable. (Papineau, 2002, ch 2.) Both Kripke's argument and the two-dimensional argument are directed solely against type=B physicalism. For this terminology of 'type-A' and 'type-B' physicalists, see Chalmers 2003.

[4[4]] I borrow this term from Bealer 2002.

which stuff actually plays this 'water role'. Similarly, if the reference of heat is fixed as the cause of heat sensations, then heat will be semantically unstable.

The crucial two-dimensional premise is then that a posteriori necessity is always due to semantic instability. (I shall call this the 'a posteriority implies instability' premise henceforth.) The thought is that, if all terms were semantically stable, then all necessities would be a priori. However, sometimes we refer to things at second hand, as it were, by using a semantically unstable concept which only hooks onto its referent with the help of the actual facts. And then it may not be a priori apparent to us that that a claim involving that concept is necessary, for we may be ignorant of the relevant actual facts and so not know what the concept refers to. Thus with water = H_2O or heat = molecular motion. Since somebody can possess the semantically unstable concepts water or heat without knowing what scientific kinds they refer to, empirical information is needed to establish that these claims are true. By contrast, the advocates of the two-dimensional argument insist, anybody who grasps a necessary claim formulated entirely in terms of semantically stable concepts will grasp the nature of the entities referred to in such a way as to render the claim a priori.

The other premise in the two-dimensional argument is that the phenomenal concept pain is not semantically unstable. Recall how Kripke argued that the appearance of pain cannot be pulled apart from its nature. In line with this, there doesn't seem any room for the phenomenal concept pain to pick out different entities, depending on the actual facts. Pain doesn't work like heat, picking out whichever entity turns out to be responsible for a certain appearance. The appearance is the pain. The phenomenal concept of pain will thus refer to the same entity—the hurty feeling—however the actual facts pan out.

This now rules out the physicalist view that pain = C-fibres firing is an a posteriori necessity. Since all the concepts involved are semantically stable[5[5]], this claim ought to be a priori if it is true. (Since the phenomenal concept of pain acquaints us directly with its referent, we ought to be able to see straight off that this is C-fibre firings, if it is.) But the claim is not a priori. So it can't be true.

4 The Physicalist Response to the Two-Dimensional Argument

In response to this argument, physicalists typically deny the 'a posteriority implies instability' premise. There seems little room to dispute that phenomenal concepts are semantically stable, picking out their referents in a way that does not depend on the actual facts. It is not so obvious, however, that semantic stability automatically guarantees an epistemological transparency that renders all necessities a priori. Even if the phenomenal concept pain is semantically inseparable from its referent, this needn't automatically ensure that mere possession of the concept will reveal all the essential features of pain.

Two strategies are open to the physicalist at this point (Levine, 2001). The 'non-exceptionalist' strategy maintains that phenomenal concepts are by no means the only counterexamples to the thesis that a posteriority implies instability. For example,

---

[5[5]] Some readers might want to query whether terms like C-fibre firings are semantically stable. Let me assume so for the moment. I shall return to this issue in section 6.2 below.

proper name concepts like Cicero and Tully don't seem to pick out their referents variably as the entities that actually satisfy certain requirements. Yet the necessary truth that Cicero=Tully is manifestly a posteriori. Again, perceptual concepts like red (applied to the surfaces of objects) arguably pick out the same property whatever the actual facts. Yet this doesn't render a putative identity like red = reflectance profile Ψ a priori.

The alternative 'exceptionalist' strategy allows that phenomenal concepts are unique in combining semantically stability with the kind of epistemological opacity that renders necessities a posteriori. This strategy grants that all other concepts that give rise to a posteriori necessity are semantically unstable, including proper name and perceptual concepts. In response to the charge that it is ad hoc to hold that phenomenal concepts are unique in combining semantic stability with epistemological opacity, exceptionalists point out that phenomenal concepts are a very special sort of concept, which refer in a distinctively phenomenological manner, and seek on this basis to explain why phenomenal concepts should also have the further unusual feature of displaying semantic stability without epistemological transparency.

To a large extent, the difference between the exceptionalist and non-exceptionalist strategies hinges on precisely how 'semantic stability' is defined (cf. Papineau 2007 sect 4.3). If this is understood in a generous manner, then proper names and perceptual concepts will also come out as semantically stable, as required by the non-exceptionalist strategy. But there are also more restrictive ways of defining semantic stability, which arguably isolate phenomenal concepts as the only semantically stable concepts that are epistemologically opaque. Still, we need not worry about these different options here. Let me now simply assume, for the sake of the argument, that physicalists can somehow explain how phenomenal mind-brain identities are posteriori even though they contains only semantically stable concepts. For what follows, it won't matter whether this explanation views phenomenal mind-brain necessities as the sole exceptions to the 'a posteriority implies instability' thesis, or as examples of a wider range of cases.

5 Kripke's Argument is Different from the Two-Dimensional Argument

I now want to show that the physicalist response (in either form) to the two-dimensional argument is no answer at all to Kripke's argument. This is because this physicalist response is designed to explain how a semantically stable necessity can appear possibly false. But Kripke's original challenge was to explain how a semantically stable necessity can appear contingently true—that is, can seem simultaneously to be actually true yet possibly false. The physicalist response to the two-dimensional argument fails utterly to meet this latter challenge.

A good initial way to see the point is to consider how those who think ordinary proper names are semantically stable (cf non-exceptionalists) will explain how the necessarily true Cicero = Tully can appear possibly false. They will simply say that a thinker can possess both these concepts but not yet know that Cicero is Tully. To such a thinker, it will remain perfectly open that Cicero is not Tully, and to this extent will appear possibly false.

Well and good. But this is no explanation for how the identity can appear contingently true. This would require Cicero = Tully to seem simultaneously true yet possible false. But how can this be? It is no good adverting to someone who does not yet know that Cicero is Tully, for that person does not take the identity to be actually true. However, if we consider somebody who does know that Cicero = Tully, it is unclear that this person can continue to view the identity as possibly false. What is this person supposed to be thinking? That the man at issue might not have been himself? Once we raise the question, it is by no means obvious how this proper name identity can appear contingently true to somebody, as opposed to merely possibly false.

I take Kripke's argument to make exactly this point about claims like pains = C-fibre firings. His challenge is not to explain how this claim could appear possibly false even if it is necessarily true. That is the challenge that physicalists can plausibly meet simply by adverting to someone who is ignorant of the identity. Rather Kripke's challenge is to explain how pains = C-fibre firings can appear possibly false to somebody who believes that it is true. And here the physicalist answer to the two-dimensional argument is no help. It is beside the point to insist that the semantic stability of the concept pain does not ensure its epistemological transparency. That might explain how someone can be ignorant that pains are C-fibre firings, and to that extent think the identity possibly false. But it is no help in explaining how somebody who fully believes that pains are C-fibre firings can continue to think that the identity might be false. Such a thinker ought to find no remaining room for the idea that pains are no C-fibre firings. It ought to be just as it is with Cicero and Tully. Just as there seems no remaining sense to the idea that Cicero might not have been Tully, once we accept that he is, so there ought to be no remaining sense to the idea that pains are not C-fibres, once we accept that they are. What are we supposed then to be thinking? That the state at issue might not have been itself?

But it does continue to appear to most of us that pains might not have been C-fibre firings, even to most of us physicalists who say that we are committed to the identity. And to this extent the case is different from the Cicero-Tully case, where it does seem that someone who fully accepts that Cicero = Tully will cease to have any room for the possibility that they are distinct. This is precisely Kripke's point. If physicalism were a tenable position, then the appearance of possible non-identity ought to disappear with the acceptance of the identity, just as in the Cicero-Tully case. But it doesn't. So physicalism isn't a tenable position.

Some readers might be worrying about my use of the Cicero-Tully analogy, and in particular about my implicit reliance on the 'non-exceptionalist' thought that ordinary proper names are semantically stable. But I hope it is clear that this assumption is needed for illustrative purposes only. Suppose, as on the 'exceptionalist view', that ordinary proper names do pick out their referents indirectly, via a priori associated descriptions. If this is so, then no doubt even thinkers who accept that Cicero is actually Tully will be able to make some sense of the thought that he might not have been Tully, for they will presumably still recognize the genuine possibility that (say) the greatest Roman orator might not have been the greatest statesman. But of course this line of thought is no help to physicalists who wants to explain why phenomenal mind-brain identities will continue to appear possibly false even to thinkers who believe them. For it is agreed on all sides that phenomenal concepts like pain are

semantically stable, however it is with ordinary proper names. So there is no question of any descriptive-style content to the concept pain, by which a thinker who does believe that pain is C-fibres firing might nevertheless construct an 'epistemic counterpart' to pain (the hurty appearance, say) which is indeed possibly different from C-fibres firing. Once you believe that pains are C-fibre firings, then there is no room for any further thought that some pain-appearing state might not be C-fibres firing. That's just the thought that pains might not be C-fibres firing, which the belief in the identity rules out. (So, even if my earlier 'non-exceptionalist' Cicero-Tully illustration of this point is unfaithful to the structure of proper name concepts, it is faithful to the structure of phenomenal concepts.)

By and large, the contemporary literature assumes that Kripke's argument is the same as the two-dimensional argument. References to Kripke's argument characteristically start by noting that it hinges on the assumption that pain is semantically stable, in that it lacks any contingent reference-fixer. This is of course entirely accurate, and constitutes one respect in which Kripke's argument matches the two-dimensional argument. But it then widely taken for granted that Kripke's argument must also share the other premise of the two-dimensional argument, the 'a posteriority implies instability' thesis.[6[6]] But this seems quite wrong. I very much doubt that Kripke thinks that any necessary claim containing only semantically stable terms must be a priori. He isn't challenging the physicalist to explain how pain = C fibres firing can appear possibly false, even if true. That's the relatively easy challenge, to which the physicalist response to the two-dimensional argument provides an answer (just think how pain = C fibres firing will appear to somebody who doesn't yet believe it). Rather Kripke's challenge is to explain how pain = C fibres firing can appear possibly false to someone who does believe it. This challenge is much harder, and it is not answered by the physicalist response to the two-dimensional argument.

So, on my diagnosis, Kripke does not embrace the 'a posteriority implies instability' premise, according to which any necessary claim that is a posteriori must contain semantically unstable terms. Instead he adopts what we might call a 'persistent possible falsity implies instability' premise: any necessary claim which continues to appear possibly false after it is believed must contain semantically unstable terms. I think that this thesis is true, and shall defend it further in the next section. But first let

---

[6[6]] Thus Brian Loar 1990 asserts that Kripke's argument shares the 'same implicit assumption' as Jackson's Mary argument, namely, 'The only way to account for the a posteriori status of a true property identity is this: one of the terms expresses a contingent mode of presentation.' David Chalmers 1996 says that Kripke's argument relies on an 'implicit endorsement of the two-dimensional framework'. Christopher Hill 1997 presents Kripke as assuming that conceivable distinctness implies real distinctness unless the commonsensical kind at issue is associated with 'a property that normally guides us in recognizing instances of X, but that is only contingently connected with it.' Joseph Levine 2001 presents Kripke as assuming that conceptual possibility implies metaphysical possibility save in cases where the claim at issue can be reinterpreted in terms of some property we use to pick out the relevant kind. Papineau 2002 attributes to Kripke the 'transparency thesis' that necessary identities will be a priori unless one of the terms refers by contingent description. Stephen Yablo 2000 goes so far as to give the name 'Textbook Kripkeanism' to the view we can move from conceivable possibility to metaphysical possibility in cases where 'no obfuscating presentation can be found' (though he himself says of this view, 'How well it corresponds to any actual belief of Kripke's is hard to say, and something I take no stand on').

me make a couple of exegetical points in defence of my understanding of Kripke's argument.

First, if Kripke had been assuming the 'a posteriority implies instability' premise, we might have expected him to articulate and defend it. In particular, we might have expected him to explain why it is not undermined by the most obvious possible counterexamples, namely, proper name identities. Since most proper name identities are manifestly a posteriori, a defender of the 'a posteriority implies instability' premise must maintain that proper names are semantically unstable. But this would be alien to the general thrust of Naming and Necessity. A central theme of the book is that there are no canonical descriptions associated a priori with ordinary proper names and that their reference is therefore fixed causally, not by description. If Kripke had thought that, even so, proper name concepts are semantically unstable, in the sense that they have a kind of content which yields different referents depending on how the actual facts turn out, we might have expected him to say so. But, as I said, there is no suggestion of this in the book.

Of course, it is possible to hold that the semantic instability of proper names is consistent with Kripke's arguments for the causal theory of reference. Thus it might be held that Kripke's arguments leave it open that Cicero has its reference fixed by the description the causal origin of the use of the name 'Cicero'. Alternatively, it might be held that Kripke's arguments leave it open that, although there are no canonical descriptions associated with proper names, individual thinkers attach idiosyncratic descriptive contents to their proper name concepts. And it might then further be held that it is precisely these descriptive contents that explain why proper names identities are a posteriori even though necessary. As I say, all these things might be held—and indeed they are held by advocates of the two-dimensional argument. (Cf. Jackson 1998, Chalmers 2002. For critical discussion, see Byrne and Pryor 2004.) But this does not show that Kripke himself held these doctrines. On the contrary, the fact that Kripke does not explicitly develop any such thoughts is surely reason to suppose that he was not thinking along these lines, and therefore that he did not embrace the 'a posteriority implies instability' thesis.

Moreover, there is some definite textual evidence that he rejects this thesis. In the pages immediately preceding his anti-materialist argument, Kripke is concerned to defend his view that a posteriori claims can be necessary, even in cases where it seems that they might have 'turned out otherwise' (p140-4). Couldn't Hesperus have turned out not to be Phosphorus? Kripke explains that strictly speaking Hesperus could not have turned out not to be Phosphorus. The only possibility around is that of a 'qualitatively identical epistemic situation' in which heavenly bodies appear in the morning and evening, as in our world, yet are different. He generalizes the point as follows:

> Any necessary truth, whether a priori or a posteriori, could not have turned out otherwise. In the case of some necessary a posteriori truths, however, we can say that under appropriate qualitatively identical evidential situations, an appropriate corresponding qualitative statement might have been false (p 142).

The crucial word here is 'some'. If Kripke thought that a posteriori necessity always involved semantic instability, then he would surely have said that such a

corresponding qualitative contingency would be available in all cases, not just 'some'. The clear implication is that he is thinking of examples like Hesperus and Phosphorus as special among proper names, in having their references fixed by salient descriptions, and correspondingly that identities involving other names will be counterexamples to the thesis that posteriori necessity implies semantic stability.

A second point in favour of my reading of Kripke is that he explicitly and repeatedly presents his anti-materialist argument as a problem for 'the identity theorist' who believes some identity claim like pain = C fibres firing. His objection is that such theorists will have no explanation of why this identity will still strike them as possibly false. ('Once again, the identity theorist cannot admit the possibility cheerfully and proceed from there; consistency, and the principle of the necessity of identities using rigid designators, disallows any such course' p 146; 'Now I do not think it likely that the identity theorist will succeed in such an endeavour' p 150; '. . . the usual moves and analogies are not available to solve the problems of the identity theorist . . .' p 155.) This is just what we should expect if Kripke's operative premise is 'persistent possible falsity implies instability', as I claim. After all, this premise bears specifically on those who believe phenomenal mind-brain identities. On the other hand, if Kripke were arguing on the basis of the 'a posteriority implies instability' premise, then we would have expected him to ask how anybody could so much as be ignorant that pains = C-fibre firings, if it is indeed true. For this latter premise, together with the semantic stability of phenomenal concepts, implies that any mind-brain identity must be a priori, if true. But Kripke doesn't query the possibility of ignorance at all. Rather he addresses his argument explicitly to 'the identity theorist' who believes some mind-brain identity claim, and then asks for an explanation of why it still seems to this opponent that this identity might have been false.

Kripke's argument, as I am reading it, thus specifically challenges physicalists to explain why mind-brain identities continue to appear possibly false to them, given that this appearance of possibility should disappear once those identities are believed. Given that the argument is framed in this ad hominem manner, one possible move open to physicalists is to explain the persistent appearance of possible falsity by saying that at some level they don't fully believe any mind-brain identities, and that this is why they appear possibly false. This seems to me the only appropriate physicalist response to Kripke. Physicalists can still hold that at a theoretical level the evidence for a range of claims like pains = C-fibre firings is compelling and sufficient for belief. But at the same time they can allow that this evidence does not give rise to an intuitive commitment to these identities, and that the persistent appearance of possible falsity is simply an upshot of the intuitive feeling that these beliefs are actually false.

In my final section I shall address the worry that this response to Kripke's argument is illegitimately ad hoc. (If we can make this response to Kripke, then why can't any reductio be blocked simply by abandoning the premise of the reductio at an intuitive level, but insisting that it is still theoretically warranted?) But first, in the next section, I shall consider whether this is really the only response that physicalists can make to Kripke.

6 Other Responses to Kripke

In effect, Kripke's argument aims to show that, when phenomenal concepts are at issue, there is no psychological gap between possible falsity and actual falsity: if a phenomenal identity claim strikes you as possibly false, then you must think that it is actually false. Given that we never think of phenomenal pain at second hand, as it were, there is no other way to breathe life into the thought that pain might not have been C-fibres firing, short of believing that they are actually different states.

I infer that, to the extent that it does seem to us that C-fibre firings might not have been pains, we do indeed believe that they are different states. We believe this at an intuitive level. We may be persuaded at a theoretical level that C-fibre firings and pains are one and the same, and so that there is no possibility of C-fibre firings without pain. But intuitively we find ourselves unable to embrace this identity, and so are intuitively unable to dismiss the possibility of C-fibres firing without pains.

In this section I want to consider whether any alternative response to Kripke's argument can avoid positing such a persistent intuition of distinctness.

6.1 Nagel's Footnote

Sometimes it is suggested that the way to explain the 'appearance of contingency'[7[7]] is to note that phenomenal imagination is quite different from perceptual imagination. This suggestion goes back to a much-cited footnote in Nagel's 'What is it Like to be a Bat?' (Nagel 1974, footnote 11. See also Hill 1997, Hill and McLaughlin 1999.) In his footnote Nagel contrasts perceptual imagination with what he calls 'sympathetic' imagination (this would now be called 'phenomenal' imagination):

> To imagine something perceptually, we put ourselves in a conscious state resembling the state we would be in if we perceived that thing. To imagine something sympathetically, we put ourselves in a conscious state resembling the thing itself. (This method can be used to imagine mental events and states—our own or another's.)

Nagel then continues:

> When we try to imagine a mental state occurring without its associated brain state, we first sympathetically imagine the occurrence of the mental state: that is, we put ourselves in a state that resembles it mentally. At the same time, we attempt to perceptually imagine the non-occurrence of the associated physical state, by putting ourselves into another state unconnected with the first: one resembling that which we would be in if we perceived the non-occurrence of the physical state. Where the imagination of physical features is perceptual and the imagination of mental features in sympathetic, it appears that we can imagine any experience occurring without its associated brain state and vice

[7[7]] Note that, if my analysis is right, the basic intuition that C-fibres could come apart from pains is not an 'appearance of contingency' at all, but simply an 'appearance of actual falsity'. Our intuition isn't that C-fibres and pains are actually identical but might have been different, but simply that they <u>are</u> different. Still, I am now discussing views that disagree with this analysis, and do take the basic intuition to be an appearance of contingency.

> versa. The relation between them will appear contingent even if it is in fact necessary, because of the independence of the disparate types of imagination.

I don't think that this works. I don't deny the feasibility of the relevant imaginative exercise, where we imagine a situation phenomenally and simultaneously perceptually imagine the absence of the corresponding brain state. But I see no reason to accept that such an imaginative exercise will give rise to an impression that the relevant situation is possible, in somebody who fully believes in the relevant mind-brain identity.

Consider, by way of analogy, somebody who fully believes that Cicero is Tully. This person can still posit someone of whom they affirm their Cicero concept and deny their Tully concept. That is, they can form the thought that someone is Cicero but not Tully. If their Cicero and Tully concepts are a priori distinct, this will be a perfectly cogent thought, free of any conceptual contradiction. But I don't see that their ability to form this though will make them feel that it is any sense possible that Cicero and Tully are distinct. They fully believe that Cicero is Tully, and so will make no substantial sense of the thought that he might not have been himself. From their point of view, the idea that someone is Cicero but not Tully will be nothing more than the empty rehearsal of a sequence of concepts, free of any conceptual contradiction, but no more pointing to a real possibility than the thought that something is both square and triangular. (Of course, there are such real possibilities as that the greatest Roman orator might not have been the greatest statesman, and the thought that someone is Cicero but not Tully may well call these to mind. But, as before, this model is no good at all to someone who wants to account for the impression that pains = C-fibre firings might have been false, since it is agreed on all sides that phenomenal concepts like pain lack the kind of structure which might allow them to be understood as referring to some epistemic counterpart.)[8[8]]

It might be objected that this analogy misses the point. It is specifically the imaginative ability to deploy phenomenal concepts in posited situations that accounts for the appearance of possibility. Maybe the mere symbolic rehearsal of Cicero but not Tully creates no appearance of possibility. But this doesn't show we won't get such an appearance when we actively imagine C-fibre firings without pains, or vice versa.

I don't see that the imaginative dimension makes any difference. Suppose I grow up listening to Elvis records, and so conceive of Elvis as the possessor of a distinctive voice. At the same time, I am familiar with visual images of a blowsy personage I think of as Presley, not realizing he is the same man. Then I discover that Elvis is indeed Presley. At this stage I will still be able to imagine a situation in which, so to speak, Elvis (and here I aurally imagine the voice) is not Presley (and here I visualise some body other than the one I think of as Presley). But will this in any sense make me feel that it is possible that Elvis might not have been Presley? I say not. If I fully

---

[8[8]] It is widely assumed that conceivability yields at least an appearance of possibility. We can now see that this is a mistake. If you retain distinct Cicero and Tully concepts, even after you come to believe the identity, you will still be able to conceive that Cicero ≠ Tully, but this non-identity will no longer appear at all possible to you (even if some epistemic counterpart does).

believe that Elvis is Presley, then I will have no room for the thought he might not have been himself, however much visualizing of his voice without his body I go in for.

Once more, there is the real possibility that someone might have sounded like Elvis without looking like him (Elvis might have sounded like Elvis without looking so blowsy). And no doubt this real possibility is called to mind by perceptually imagining Elvis's voice attached to a different body. But as before this model is no good for explaining how mind-brain identities appear possibly false, given that there is no question of separating the impression created by pains from the pains themselves.[9][9]

6.2 Semantic Instability on the Right-Hand Side

So far I have assumed that the physical concepts involved in mind-brain identity claims are semantically stable, along with the phenomenal concepts. That is, I have assumed that the reference of concepts like C-fibre firings does not depend on how the actual facts turn out.

However, there is plenty of room to doubt this assumption of semantic stability. It is a familiar thought that theoretical terms in science have their references fixed by description. On this account, theoretical entities are picked out as those items that bear such-and-such causal relations to measuring instruments and human observers and to other theoretical entities.[10][10] If this view of scientific terms is right, then they will not be semantically stable. For their referential value will vary with the actual facts, depending on which entities actually play the causal role in question. Applying this model, the concept C-fibre firings thus comes out as similarly semantically unstable—some actual arrangement of basic entities will play the C-fibre role in this world, but something different might have played this role if the actual world had turned out to be constituted by different basic entities.

This now suggests an alternative explanation for the appearance of contingency associated with mind-brain identity claims. Perhaps such claims strike us as possibly false, not because we intuitively disbelieve them, but because we have in mind that something other than the actual realizer might have turned out to play the relevant scientific role. This would be akin to the standard explanation of why water = H_2O and heat = molecular motion strike us as apparently contingent—except that now we will be considering a possible dissociation between role and actual realizer on the scientific right-hand side (C fibres firing) rather than on the commonsensical left-hand side (pain).

---

[9][9] Interestingly, Christopher Hill 1997 agrees with this diagnosis. Hill uses Nagel's distinction between sympathetic and perceptual imagination to explain how it can intuitively appear to us that C-fibre firings are separable from pain, even if they are in truth identical. But he also says (in his own footnote 11) that such intuitions can be defeated if we have reasons for believing that they are necessarily false. From my perspective, Hill thus fails to address Kripke's actual argument, as opposed to the two-dimensional argument. On my reading, Kripke's argument hinges precisely on the fact that intuitions of separability are not defeasible in the way Hill supposes.

[10][10] For any given theoretical concept T, this account can be formalized by positing a 'Carnap sentence' which says that If there are any entities which play the relevant causal role, then they are Ts, and regarding this sentence as implicitly defining T.

So the suggestion is that the 'appearance of contingency' associated with mind-brain claims like pain = C fibre firings arises because we are aware of the possibility that that something other than the actual realizer might play the C-fibre role. In the actual world, this role is played by some specific basic state, and this basic state is identical to pain. But some different basic state—some state other than pain—might have played the C-fibre role, and that is what we are thinking about when we think that there could be C-fibre firings without pain. Or so at least this suggestion goes.[11[11]]

I find this suggestion unpersuasive. I have no objection to the idea that theoretical terms in science refer by description, nor therefore to the thought that they might have turned out to be realized by something different from their actual realizers. Accordingly, I am happy to agree that it is genuinely possible that the C-fibre role might have turned out to be realized by some other basic state than its actual realizer—that is, by something other than pain—and I therefore recognize that this is a cogent way of giving substance to a judgement that C-fibre firings might not have been pains, even though they actually are. However, I don't accept that this is what we are ordinarily thinking when we think that there could be C-fibre firings without pain.

The suggestion is surely far too complicated. When we ordinarily think of C-fibres without pains, we don't consider alternative realizers for the C-fibres role. Rather, we simply hold fixed the C-fibres and whatever realizes them—we keep everything the same in the brain, so to speak—and then judge that, even so, pain could be absent. Our thought is that everything relevant to the presence of C-fibres could be just as it is in the actual world, and yet there be no pains.

Thoughts about alternative realizers for the C-fibre role are thus the wrong 'shape' to explain our intuitive conviction that pains could come apart from C-fibre firings. They may point to genuine possibilities, but they achieve this only by supposing some difference in the way C-fibres are realized. They are thus no good for explaining the intuition that C-fibres could be realized just as they actually are, and yet there be no pains.

6.3 Mightn't We Fail to Realize That Identities are Necessary?

Aren't I crediting ordinary thinkers with a great deal of sophistication, when I say that, if they really believe any phenomenal mind-brain identities, then they will have no remaining room for the thought that these identities might be false? Doesn't this presuppose that they understand that identity is a necessary relation? But this is a subtle matter, which not everybody appreciates clearly, certain not those who haven't had the benefit of Kripke's first two Lectures in Naming and Necessity.

[11[11]] The two-dimensional 'a posteriori necessity implies semantic instability' thesis implies that, if we could ever refer to the realizer of C-fibre firings in a semantically stable way, then it would become a priori that the realizer is pain; some philosophers take this to motivate some kind of 'neutral monism' (cf. the discussion of 'type-F monism' in Chalmers 2003). These thoughts go beyond the suggestion currently being considered, which requires only that the C-fibre role has a contingent actual realizer, and assumes nothing about the epistemic transparency of alternative ways of referring to this realizer.

This then suggests a different explanation for why many physicalists think that it is possible for C-fibre firings not to be pains. It's not that they don't fully believe the identity. It's simply that they don't appreciate that identities are necessary.[12[12]]

As it happens, I don't think that I am crediting ordinary thinkers with too much sophistication, in supposing that they appreciate the necessity of identity. At bottom, I'm only assuming that, if they think that entity X is identical to entity Y, then they can't think that it—that entity—might not have been itself. And that seems pretty basic to me.

But I am happy to let this point pass. For I can instead rest my case on those clear-headed physicalists who have read Kripke and who do clearly grasp that identities could not have been otherwise. I say that even they (us) will have an impression that C-fibre firings could occur without pains, even after being shown all the positive evidence for their identity. But now, by hypothesis, this impression of possible falsity can't be due to lack of appreciation of the necessity of identity. So the only explanation, once more, is that even we sophisticates find it impossible fully to believe our physicalism.

And if even we sophisticates find it impossible to free ourselves of an intuition of dualism, then that is surely every reason to suppose that ordinary thinkers have that intuition as well.

6.4 Mightn't Pains have 'Turned Out Not To Be' C-Fibres Firing?

If we did come fully to believe a mind-brain identity claim like pain = C-fibres firing, wouldn't we still have room for the thought that pains might not have turned out not to be C-fibres firing? And mightn't this epistemological possibility alone account for the appearance of contingency, even after we accept that it is metaphysically necessary that pains are in fact C-fibres firing? Surely our becoming fully convinced that pain = C-fibres firing needn't stop us thinking that the evidence might have turned out differently, and that in this epistemological sense pains might not (have turned out to) be C-fibres firing after all?

But this does not work. If the analysis of this paper so far is correct, then fully believing that pain is C-fibres firing will destroy any epistemological possibility of its being different, along with any metaphysical possibility thereof. The pain case isn't like water = H_2O, say. In the latter case, the discovery that water is H_2O does indeed leave a sense in which it might have turned out not to be. This is because the identity of water with H_2O does not rule out possible worlds where the watery stuff is XYZ, say. So we can continue to recognize the genuine possibility that we might have been in one of those worlds, worlds we would have discovered that the watery stuff is XYZ. And it is natural enough to describe this as the possibility that 'water' (ie the watery stuff) might have turned out not to be H_2O

But with pain there are no 'epistemic counterparts', nothing which stands to pain as watery stuff stands to water. A world in which the 'hurty stuff' is not C-fibres firing

---

[12[12]] Christopher Hill explicitly offers this as an explanation of why even someone who is fully informed of a scientific identity can think it is possibly false (1997, footnote 14).

is a world in which pain is not C-fibres firing, and so is ruled out by the knowledge that pain is that process. So, if you fully embrace the claim that pain = C-fibres firing, you will therewith cease to allow any sense in which pain—that very feeling—might have turned out to be something else. If it is C-fibres firing, how could it have been something else?

Of course, even after you accept that pain = C-fibres firing, you can still allow that the scientists might have announced that pain is something else. Indeed you might suppose that such a thing will come to pass in the actual future. But, as long as you adhere to the claim that pain is C-fibres firing, you must think that any such scenario would involve some kind of mistake on the part of the scientists. Perhaps it is a scenario in which they are using the word 'pain' to refer to some different conscious state. Perhaps they have been misled by some freak evidence. What is not a possibility is that they might correctly conclude pain is not C-fibres firing. (Suppose you now firmly believe Fermat's last theorem. You can still think it possible that Andrew Wiles might one day call a press conference and say that it is not true after all. But you can't allow that he might be correct in so saying, as long as you continue to believe the theorem.)

6.5 Do Physicalists Believe Any Mind-Brain Identities?

So far I have been proceeding on the assumption that physicalists are people who fully believe some specific mind-brain identities like pain = C-fibres firing. But in fact this does not accurately describe contemporary physicalism. This is because neurophysiology is as yet too underdeveloped to support any specific mind-brain identity claims. After all, it is not an accident that philosophers always use the silly example of pain = C-fibres firing, even though we know that this is not a good account of pain. The reason is that we don't yet have any clear-cut examples of well-established equivalences between specific conscious states and specific physical ones.

This now suggests a straightforward explanation for our intuitive impression that it is possible to have C-fibres firing without pains, and similarly for any other suggested pairing of brain state with phenomenal state. Perhaps these impressions arise simply because we don't in fact believe the relevant mind-brain identities, and so attach a positive credence to their actual falsity. We think that there is some positive probability that any such identity is actually false, and a fortiori that there is some positive probability that it is possibly false.

I agree that this story offer an adequate explain why it should seem possible to us that there could be C-fibres without pains, and similarly why it seems to us that other specific mind-brain associations might come apart.

Still, such specific dissociations aren't the only way in which physicalism strikes us as possibly false. We also have the more general impression that zombies are possible, in the sense that it strikes us that there could be beings that shared all our physical states but yet had no phenomenal states. And this impression is ruled out with something that we do have every reason to believe, despite the underdeveloped state of neuroscience, namely, the general physicalist thesis that every conscious states is identical to some physical state or other.

I take this alone to establish that we are in the grip of an intuitive resistance to physicalism: in order to explain why we still think that zombies are possible, despite the evidence for a general physicalism, we need to recognize that something is stopping us fully embracing this physicalism.

Let me go more slowly. My argument here involves two claims: first, we already have reason to believe generalized physicalism, even in the absence of evidence for specific mind-brain identities; second, zombies should cease to appear possible to anybody who whole-heartedly believes this generalized physicalism. Let me take these in reverse order.

To see why fully believing generalized physicalism should eliminate the impression that zombies are possible, note that the zombie thought is stronger than the thought that someone could have C-fibres without pain (or P' without pain, or P'' . . ., for some specified list of physical states). Rather it is the thought that someone could have <u>all</u> our physical states (whatever they are) and yet not have pain (or any other conscious state). And this stronger thought is surely ruled out by even the relatively weak general physicalist claim that every conscious state is identical to <u>some</u> physical state. After all, this general claim obviously implies that your physical duplicate will have all the physical states, whatever they are, that are identical with your conscious states, and so will have those conscious states too.

Consider an analogy. You are told that Cary Grant is identical to one of twenty named people, but you aren't told which one. Will you think it possible that those twenty people are in a room, yet Cary Grant not be there? Not if you fully believe that Cary Grant is one of those twenty. Similarly the thought that zombies are possible would be eliminated, if we unequivocally believed generalized physicalism.

My other claim is that we already have reason to believe generalized physicalism, even without evidence for specific mind-brain identities. Recall how I said at the beginning of this paper that 'the positive case for physicalism is familiar'. The familiar case I had in mind was the 'causal argument' that starts with the causal completeness of the physical realm (every physical effect has a fully physical causal history) and concludes that that conscious states must themselves be part of the physical realm (otherwise they will be 'causal danglers' which never make any difference to what happens in the physical realm). I take this to be a powerful argument for the general conclusion that every conscious state must be identical to <u>some</u> physical state, even though it tells us nothing about what specific physical states these are.

True, there are philosophers who aim to evade even this general physicalist conclusion, perhaps by denying the causal completeness of physics, or alternatively by accepting that conscious states are indeed epiphenomenal. But we do not need to settle this debate here. It is enough for present purposes to note that there are plenty of philosophers (myself for one) who regard the causal argument as mounting a conclusive case for general physicalism. Even so, zombies still strike me and many other professed physicalists as intuitively possible. Given that a whole-hearted commitment to a general physicalism would destroy any such impression, it follows that something is stopping us self-proclaimed physicalists from properly believing our general physicalism.

It remains possible that, even so, future scientific developments will somehow strengthen our commitment to physicalism to such an extent that we will cease to regard zombies as possible. Stephen Yablo thinks so. 'Am I the only one who feels the intuition of zombies to be vulnerable in this way?' he asks (2000, p 119).[13][13] Perhaps Yablo is right, and in time the intuition of distinctness will fade away. Still, as things presently stand, something is preventing most of us from fully embracing physicalism, even those of us who can see at a theoretical level that it must be true. To know whether this intuitive barrier to fully accepting materialism will dissolve as a result of future scientific developments would require a more detailed understanding of the nature and origin of this barrier than I have attempted in this paper.

For what it is worth, though, I suspect that the intuition of mind-brain distinctness is here to stay. In this respect, I take it to be similar to the many other familiar cases where human intuition persistently continues to reject something that we know to be true at a theoretical level. (Cf. Weatherson, 2003.)

Thus consider knowingly experienced perceptual illusions, like the familiar Muller-Lyer lines. No amount of theoretical knowledge makes these illusions disappear. We can know full well that the lines in the Muller-Lyer illusion are the same length, but they persist in looking different lengths. Nor need this kind of set-up always involve a conflict is between a perceptual and non-perceptual judgement. There are examples where both judgements are non-perceptual. Thus consider the theory that there is no moving present, and that 'now' has a purely indexical semantics. Philosophical analysis has convinced me that this is true, but in my heart I cannot really believe that my present life is ontologically on a par with that I had forty years ago. I understand general relativity well enough to know that time does not extend before the big bang, but at an intuitive level that doesn't stop me wondering what went on before. I am convinced that Everett's no-collapse theory is the only remotely plausible interpretation of quantum mechanics, but this doesn't stop me thinking that the cat will either be alive or dead when I open Schrödinger's box. And so on.

A proper treatment of these cases would require an explanation of how our cognitive system can segment itself into an 'intuitive' and a 'theoretical' part, and thereby stably contain two contradictory judgements. But I hope these examples will at least persuade you that this is possible. My view is that this is how it is with physicalism. At a theoretical level we may be sure that physicalism is true, and so that zombies are quite impossible. But there is something in the way we think of conscious states that inevitably makes us feel that phenomenal states are distinct from brain states, even if perfectly correlated in actuality, and therewith that zombies are possible after all.

7 Psychology in, Psychology Out

In this final section I want to address the worry that my response to Kripke is ad hoc.

Kripke points out that a commitment to physicalism is incompatible with the 'appearance of mind-brain contingency'. I accept this, responding that the appearance

---

[13][13] Hill 1997 is also implicitly committed to the thesis that zombies will cease to appear possible to anybody who both believes physicalism and appreciates the necessity of identity.

arises only because we physicalists are less that fully committed to our physicalism at an intuitive level. Nevertheless, I insist, our physicalism is entirely acceptable at a theoretical level.

Some readers may feel that this is a cheap move. If I can respond to Kripke like this, why shouldn't anybody be able to block any reductio argument similarly? Somebody points out that your view p implies some unacceptable q. You agree that q is unacceptable, and that p implies it, but maintain that the difficulty can be dealt with by noting that you disbelieve p at an intuitive level. Nevertheless, you insist, there are good arguments for p, and you will continue to embrace it at a theoretical level.

Clearly this is no good. If p implies some falsehood, then p is false, and that's the end of it. But where then does my response to Kripke differ?

The crucial point to note is that Kripke isn't offering a straightforward reductio of physicalism. Physicalism implies that zombies are impossible. If Kripke could show that zombies are possible, then that would indeed directly refute physicalism. But Kripke's argument does not assume that zombies are possible. He realizes that this would beg the question. ('. . . the identity theorist is committed to the view that could not be a C-fiber stimulation which was not a pain nor a pain which was not a C-fiber stimulation. These consequences are certainly surprising and counterintuitive, but let us not dismiss the identity theorist too quickly' p 149).

Rather, Kripke's argument assumes, not that zombies are possible, but only that they <u>appear</u> possible. This is a psychological premise, not a metaphysical one. Insofar as Kripke is attempting a reductio, it must therefore go like this: physicalism implies that zombies will not even appear possible; but zombies do appear possible; so physicalism is false.

Putting it like this, it becomes clear that my appeal to an intuition of dualism isn't a misguided attempt to side-step a sound reductio argument. Rather, I am simply disputing the first premise in the reconstructed reductio, that is, the claim that physicalism implies that zombies won't even appear possible. We can think of this premise as a challenge to physicalists to find some explanation for the apparent possibility of zombies. Kripke points out that the kind of explanation that we might offer for the apparent possibility of heat ≠ molecular motion will not work in the mind-brain case. But he himself allows that some other explanation might be possible. ('I certainly cannot discuss all the possibilities here' p 155.)

I have here offered just such an alternative explanation. The psychological fact that zombies appear possible is a consequence of another psychological fact—phenomenal and physical states strike us as intuitively distinct. In the context of Kripke's argument, there is nothing ad hoc about this move. In effect, Kripke asks for an explanation of a psychological fact. It is entirely appropriate to offer another psychological fact in response to this request. Moreover, it is clear that there is no incompatibility between this psychological fact and the metaphysical thesis of

physicalism itself. Many things that strike human beings as intuitively false nevertheless turn out to be true.[14[14]]

Bibliography

Bealer, G. 2002 'Modal Epistemology and the Rationalist Renaissance' in Hawthorne, J. and Gendler, T. (eds) Conceivability and Possibility Oxford: Oxford University Press

Block, N. and Stalnaker, R. 2000 'Conceptual Analysis, Dualism and the Explanatory Gap' Philosophical Review108.

Byrne, A. and Pryor, J. 2004 'Bad Intensions' in Garcia-Carpintero, M. and Macia, J. (eds) The Two-Dimensional Framework: Foundations and Applications Oxford: Oxford University Press

Chalmers, D. 1996 The Conscious Mind Oxford: Oxford University Press

Chalmers 2002 'On Sense and Intension' Philosophical Perspectives 16

Chalmers, D. 2003 'Consciousness and its Place in Nature' in Stich, S. and Warfield, F. (eds) The Blackwell Guide to Philosophy of Mind Oxford: Blackwell

Chalmers, D. 2006 'The Two-Dimensional Argument Against Materialism' in his The Character of Consciousness Oxford: Oxford University Press

Hill, C. 1997 'Imaginability, Conceivability, Possibility and the Mind-Body Problem' Philosophical Studies 87

Huxley, T.H. 1866 Lessons in Elementary Physiology London: Macmillan

Jackson, F. 1993 'Armchair Metaphysics' in O'Leary-Hawthorne, J. and Michael, M. (eds) Philosophy in Mind Dordrecht: Kluwer

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

Jackson, F. 1998 'Reference and Description Revisited' Philosophical Perspectives 12

Kripke, S. 1980 Naming and Necessity Oxford: Blackwell

Levine, J. 1983 'Materialism and Qualia: The Explanatory Gap' Pacific Philosophical Quarterly 64

Levine, J. 2001 Purple Haze New York: Oxford University Press

[14[14]] Versions of this paper have been delivered to seminars at Birmingham, Cambridge, North Carolina at Chapel Hill, Sussex, Berlin, and King's College London. I would like to thank all those who made comments on those occasions, especially David Chalmers, Keith Hossack, Christina Nimtz, Mark Sainsbury and Gabriel Segal

Loar, B. 1990 ‘Phenomenal States’ in Tomberlin, J. (ed.) Philosophical Perspectives 4

Melnyk, A. 2003 ‘Papineau on the Intuition of Distinctness’ SWIF Forum on *Thinking about Consciousness* http://lgxserver.uniba.it/lei/mind/forums/004_0003.htm

Papineau, D. 1993 ‘Physicalism, Consciousness, and the Antipathetic Fallacy’ Australasian Journal of Philosophy 71

Papineau, D. 1998 ‘Mind the Gap’ in Tomberlin, J. (ed.) Philosophical Perspectives 12

Papineau, D. 2002 Thinking about Consciousness Oxford: Oxford University Press

Papineau, D. 2006 ‘Comments on Strawson’s “Realistic Monism: Why Physicalism Entails Panpsychism”’ Journal of Consciousness Studies 13

Papineau, D. 2007 ‘Phenomenal and Perceptual Concepts’ in Alter, T. and Walter, S. (eds) Phenomenal Concepts New York: Oxford University Press

Weatherson, B. 2003 ‘What Good are Counterexamples?’ Philosophical Studies 115

Yablo, S. 2000 ‘Textbook Kripkeanism and the Open Texture of Concepts’ Pacific Philosophical Quarterly 81