Philosophical Perspectives, 21, Philosophy of Mind, 2007

KRIPKE'S PROOF IS AD HOMINEM NOT TWO-DIMENSIONAL

David Papineau

1. Kripke's Argument

In the final pages of *Naming and Necessity* Kripke offers an argument against mind-brain identity theories. It runs like this.

> Identity theorists make claims like 'pain = C-fibre stimulation'. These claims must be necessary if true, given that terms like 'pain' and 'C-fibre stimulation' are rigid. Yet there is no doubt that such claims appear contingent. It certainly seems that there could have been C-fibre stimulation without pains or vice versa. So identity theorists owe us an explanation of why such claims should appear contingent if they are in fact necessary.
>
> One model for such an explanation would be the story we use to explain why scientific identity claims like 'heat = molecular motion' appear contingent. In that case we can say that this appearance of contingency is due to our recognition of the genuine possibility that molecular motion might not have been *felt as heat*. However this won't work in the mind-brain case. There is no corresponding possibility that C-fibre stimulation might not have been *felt as pain*, if pain is in fact C-fibre stimulation. Pains can't be pulled apart from their appearance, in the way that heat can. A situation in which C-fibre stimulation isn't felt as pain is a situation which lacks *pain*, not just the appearance of pain. And this possibility—of C-fibre stimulation without pain—is inconsistent with the identity theorist's insistence that pain *is* C-fibre stimulation.
>
> Given that no other explanation of the apparent contingency of 'pain = C-fibre stimulation' offers itself, identity theorists have no option but to accept that this claim is not necessary after all, and so not true.'

This paper will first focus on the correct exegesis of this argument and will then consider what substantial points it establishes.

2. The Orthodox Interpretation

There is currently a widespread consensus on the way to read this argument. According to nearly all contemporary commentators, when Kripke refers to

the 'appearance of contingency' displayed by mind-brain identity claims, he is in effect talking about the *a posteriority* of these claims. Of course, these commentators do not think that Kripke equates a posteriority with contingency itself—that would run counter to his fundamental separation of epistemology and metaphysics. But they do think that Kripke takes a posteriority to give rise to an *appearance* of contingency. (After all, the thought runs, a posteriority leaves it open whether or not a claim is true.)

Furthermore, according to this contemporary consensus, Kripke holds that such an appearance of contingency can only be explained if some term in the relevant claim picks out its referent by contingent description, or at least can plausibly be understood as having such a descriptive content. Thus, in the heat = molecular motion case, the term 'heat' can be understood as equivalent to 'the cause of our sensations of heat'. This then enables us to (mis)understand that claim 'heat = molecular motion' as 'the cause of our sensations of heat = molecular motion'. Since this latter claim is genuinely contingent, this then offers an account of the appearance of contingency associated with the former claim. However, no such descriptive reading is available in mind-brain cases, which means there is no way to explain why they too should appear contingent.

So the contemporary consensus reads Kripke as assuming that a posteriori necessities are always due to the presence of descriptive content. In support of this interpretation, we might suppose that Kripke was thinking along something like the following lines. When we think of some referent via a description, we are thinking of it at second hand, thinking of it via some associated property. So thinking of this kind may well conceal necessities from us—the indirecteness of our thought may prevent necessary properties from being a priori discernable. By contrast, if we are thinking of something directly, without the mediation of some contingent description, then any necessary properties will be available a priori. (And then, in the case of pain, where our thoughts clearly aren't mediated by any description, any necessary property of pain, such as its putative identity with C-fibre stimulation, ought to be available a priori. But pain's identity with C-fibre stimulation clearly isn't available a priori. So pain can't in fact be identical with C-fibre stimulation.)

On this reading, then, Kripke's argument hinges crucially on the principle:

(I) If a necessarily true claim is a posteriori, then at least one of its terms must be associated with a descriptive content.

Some of those who hold that Kripke is committed to (I) also take him to be implicitly committed to the doctrines of 'two-dimensional semantics' and in particular to the claim that every a posteriori necessity has an 'unstable primary intension' which is sensitive to the actual-world facts in the way that the denotations of contingent descriptions are.[1] Other philosophers hesitate to attribute any such systematic view to Kripke on the basis of the discursive and unmethodical remarks in *Naming and Necessity*, but even so feel that there is

enough in that text, and in particular in the implicit logic of the anti-materialist argument at the end of the book, to make it clear that Kripke must accept something along the lines of principle (I).

Let me cite some examples of the orthodox interpretation from the recent literature.

In his 'Phenomenal States' (1997) Brian Loar asserts that Kripke's anti-physicalist argument hinges on the assumption that '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' (p 600).

Christopher Hill (1997) puts the point in terms of inferences from conceivable distinctness to real distinctness. The conceivable distinctness of some commonsensical kind X (eg pain) from a theoretical kind (C-fibre stimulation) means that their identity can at best be a posteriori. So principle (I) would mean that such conceivably distinct kinds can only be identical if at least one is picked out descriptively. In line with this, Hill attributes to Kripke the view that such conceivable distinctness does imply 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 X' (p 63).

Again, Joseph Levine's *Purple Haze* (2001, pp 47–8) 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. And in my *Thinking about Consciousness* (2002, pp 92–3) I myself followed the trend and attributed to Kripke the 'transparency thesis' that necessary identities will be a priori unless one of the terms refers by contingent description.

In the Introduction to their anthology on two-dimensional semantics (2006) Manuel García-Carpintero and Josep Macià allow that there is no explicit commitment to anything like (I) in *Naming and Necessity*: they say that Kripke avoids any general claim that any posteriori necessities must hinge on the presence of a descriptive term that gives rise to 'an appropriate corresponding qualitative statement'. Nevertheless they argue specifically that his argument against mind-body identity 'depends essentially' on this assumption, and so that he must be implicitly committed to it (p 2). Here they are following David Chalmers line in *The Conscious Mind* (1996). Chalmers similarly maintains that Kripke's argument suggests an 'implicit endorsement of the two-dimensional framework' (p 149).

Stephen Yablo 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' (2000, p 100). It should be noted, however, that Yablo himself is non-committal about this interpretation. ('How well it corresponds to any actual belief of Kripke's is hard to say, and something I take no stand on.')

3. Can Kripke Really Be A Textbook Kripkean?

Can Kripke really be committed to principle (I)? It seems to me that such a commitment would run quite counter to the central doctrines of *Naming and Necessity*. After all, much of the book is devoted to persuading readers that most ordinary proper names lack descriptive content, but rather have their referents fixed causally. Given this, it looks as if proper name identities are prima facie counter-examples to principle (I). These necessities are manifestly a posteriori, but arguably lack any descriptive content. Yet Kripke seems quite unconcerned about this. Maybe there is more to say here—maybe the text leaves it open that Kripke thinks proper names all carry some kind of associated descriptive content alongside their causal referential ties. But the point remains that Kripke himself does not say anything more in *Naming and Necessity*, as he surely would if his anti-materialist argument did hinge on principle (I).

Further, consider the epistemological implications of principle (I). This principle says that necessary claims can only be a posteriori if some of their terms refer by description. It follows immediately that when reference is direct, and not mediated by description, then all necessary properties will be a priori—we will be acquainted directly with the objects, and so their nature will be transparent to us. I see no hint of this kind of Russellian epistemology in Kripke. He has plenty to say about reference that isn't mediated by description, but no suggestion that such reference renders all necessary properties epistemologically transparent.

I shall have more to say against the orthodox reading of Kripke's anti-materialist argument below. But first it will be helpful to show that there is an alternative way of understanding Kripke that does not commit him to Principle (I).

4. Kripke's Ad Hominem Argument

I think that Kripke's argument is best read as an ad hominem challenge to identity theorists, rather than as a manifestation of some implicit Russellian epistemology. It is instructive to see how Kripke formulates his argument From the beginning he presents the issue as a problem arising specifically for 'identity theorists'—those who embrace claims like 'pain = C-fibre stimulation'—and not as an instance of some general constraint on a posteriori necessity. Thus he says:

> ... the identity theorist is committed to the view there could not be a C-fibre stimulation which was not a pain nor a pain which was not a C-fibre stimulation .... Can he perhaps show that the apparent possibility of pain not having turned out to be C-fibre stimulation ... is an illusion? ... Now I do not think it likely that the identity theorist will succeed in such an endeavour. (pp 149–50)

There is no suggestion here that there is some general problem about posteriori necessities with no descriptive content. The problem isn't that any such necessity

ought to be a priori available to every thinker. Rather Kripke is pointing out that there is a specific issue facing thinkers who actually *believe* that pain = C-fibre stimulation. These people need to explain why this identity still strikes *them* as possibly false. People who *aren't* committed to the identity will of course countenance worlds in which pain turns out not to be C-fibre stimulation. But once you embrace the identity it ceases to be at all clear how you can still have room for the thought that pains might not have been C-fibre stimulation.

What are you supposed to be thinking, when you think this? You think that pain and C-fibre stimulation are one and the same thing. So how can you countenance a world in which 'they' come apart? Of course, if 'pain' (or 'C-fibre stimulation') had some descriptive content, then you could say that you were imagining a world in which something else satisfied the description is question, just as we can imagine a world in which molecular motion doesn't satisfy the description associated with 'temperature'. But there isn't any such descriptive content available in the pain = C-fibre stimulation case. So you really can't explain why this identity strikes you as contingent, given your contention that it is actually true.

Consider this analogy. People who do not know that Cicero is not Tully will of course think it possible that they are two people. But suppose that you come firmly to believe that Cicero *is* Tully. Can you still think that 'they' might have been two people? What would you be thinking? That *this* person might not have been himself? Surely you will no longer have any room for the thought that Cicero and Tully might have been different people.

Of course, in proper name cases there may be some surrogate contingency in the offing, associated with your attaching some descriptive content to 'Cicero' or 'Tully'. For example, if you associate the descriptions 'the greatest Roman orator' and 'greatest Roman statesman' with these names, then you could read the identity as the uncontroversially contingent claim that 'the greatest Roman orator might not have been the greatest Roman statesman'. But, once more, we won't be able to explain the apparent contingency of mind-brain identities in an analogous way, given that no such descriptive reading seems available in those cases.

On my reading, then, Kripke's challenge isn't to explain how mind-brain identities are a posteriori—as it were, to explain how they can appear possibly false to people who don't yet believe them. Rather his challenge is to explain why they *still* appear possibly false, even to people who *do* believe them. Given that there is no descriptive content to hand, there seems no room for someone simultaneously to think that pain *is* C-fibre stimulation, but that it might not have been.

So construed, Kripke's argument still hinges on the claim that 'pain' (and C-fibre stimulation') lack descriptive content, just as it does on the orthodox interpretation. But on my reading he isn't committed to Principle (I) at all. Rather his crucial premise is:

(II) If a necessary truth still seems contingent after it is believed, then it must have some descriptive content.

5. Kripke's Actual Argument Can't Be Blocked As Easily As The Orthodox Interpretation

To further bring out the difference between my reading of Kripke and the orthodox interpretation, note that the standard physicalist rebuttal of the orthodox interpretation is ineffective against Kripke as I am reading him.

On the orthodox interpretation, as we have seen, Kripke's argument is that mind-brain identities lack the descriptive content that principle (I) says is characteristic of any a posteriori identity. To this the normal physicalist response is to deny principle (I). Contemporary physicalists simply insist that there can be a posteriori identities even in the absence of any descriptive content. Sometimes two terms can refer directly and yet their co-reference not be a priori.[2]

However this is no help in answering Kripke's argument as I am construing it. Even after you deny principle (I), you will still face Kripke's ad hominem argument.

Denying principle (I) allows materialists to hold that mind-brain identities are a posteriori, and therewith to explain how these identities can appear possibly false to people who do not yet believe them. But this doesn't help to explain how such description-free identities can still appear possibly false even to people who *do* believe them. Let us agree that pain = C-fibre stimulation is a posteriori and so rests on empirical evidence. Still, how come this claim *still* appears possibly false to us, even after we have the evidence? Shouldn't this appearance simply disappear, once you accept that pain *is* C-fibre stimulation, and have no descriptive content with which to create a contingent surrogate for this claim? Kripke's challenge as I am construing it simply isn't addressed by the standard physicalist insistence that mind-brain identities can be a posteriori.

6. Further Evidence Against The Textbook Reading of Kripke.

In section 3 I made two initial exegetical points against the orthodox interpretation of Kripke:

(i) He doesn't offer any explanation of why proper name identities aren't counterexamples to principle (I), as one would expect if he endorsed that principle.
(ii) He doesn't seem to embrace the kind of Russellian epistemology—direct reference renders essential properties epistemologically transparent—that is implied by principle (I).

Let me now add to the exegetical case. For a start, note how Kripke does not object directly to the idea that mind-brain identities are a posteriori, as we might

surely expect if he were arguing on the basis of principle (I). After all, principle (I) says that mind-brain identities must be a priori if true. So why doesn't Kripke simply argue that it absurd to suppose that 'pain = C-fibre stimulation' might be known a priori? However, this isn't what he does. Rather he starts talking about the possible ways in which 'the identity theorist' might explain the 'appearance of contingency'. This strongly suggests that when he talks about the 'appearance of contingency' he is referring to something more complicated than the mere a posteriority of mind-brain identities.

It is also noteworthy that Kripke here uses the specific phrase 'appearance of *contingency*'. If the complaint were simply that mind-brain identities can't be a posteriori, then the problem would be to explain why they should appear possibly false to anybody. But Kripke doesn't ask for an explanation for the appearance of possible falsity, but more specifically for the appearance of *contingency*. I take 'contingency' in this context to mean contingent truth, that is, the combination of truth in the actual world with falsity in other possible worlds. It is this specific combination that 'the identity theorist' seems to be committed to, and which I say Kripke views as problematic. If Kripke were not arguing ad hominem against 'identity theorists' who are already committed to the truth of mind-brain identities, then it would be mysterious that he asks about the appearance of 'contingency' rather than a mere appearance of possible falsity.[3]

In addition to all these implicit considerations against the orthodox interpretation of Kripke, there is also some direct textual evidence that he rejects principle (I).

Thus, just before he starts on his anti-materialist argument, Kripke is concerned to defend the necessity of rigidly framed identity claims, against such thoughts as that Hesperus might not have turned out to be Phosphorus. Kripke explains that such 'might have turned out' thoughts hinge on reinterpreting the relevant claims in term of something like descriptive content. Thus there is a possible world in which the star seen in the morning is not the star seen in the evening, even though the actual Hesperus is necessarily the actual Phosporus. More generally, rigidly framed identity claims can often be re-read as contingent claims about objects which satisfy salient qualitative criteria. Here is how Kripke puts the point.

> 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, my italics).

I have italicized the 'some' in the second sentence. This seems to tell decisively against the view that Kripke's anti-materialist argument assumes principle (I). If he were about to embark on an argument that hinged crucially on the premise that *all* a posteriori necessities involve some kind of descriptive content,

surely he would not say, two pages earlier, that such contents are associated with 'some' a posteriori necessities. The clear implication is that Kripke takes examples like 'Hesperus' and 'Phosphorus' to be relatively special among proper names, and accepts that in other cases no 'corresponding qualitative statement' will be available. If this is right, then his argument against mind-brain identities can't possibly be that there are no brute (non-descriptive) a posteriori necessities.

There is a similar passage a few pages later, in the middle of the anti-materialist argument itself. Kripke is now considering the ways in which the 'identity theorist' might explain the 'appearance of contingency'.

> What was the strategy used above to handle the apparent contingency of certain cases of the necessary *a posteriori*? The strategy was to argue that although the statement itself is necessary, someone could, *qualitatively* speaking, be in the same epistemic situation as the original, and in such a situation a *qualitatively* analogous statement could be false (p 150).

Kripke then proceeds to explain how this strategy won't work in the mind-brain case, for lack of any suitable qualitatively analogous contingent statement.

Note how Kripke here refers to the strategy used to 'handle the apparent contingency of *certain* cases of the necessary a posteriori'. Kripke does not say that the qualitative strategy will be available in *all* cases of the necessary a posteriori. Rather the suggestion is that it ought to be available specifically in 'certain' cases—namely, those that display 'apparent contingency' by continuing to appear possibly false even after they are believed to be true. This seems to make it quite clear that Kripke's argument isn't that descriptive content is needed for a posteriori identities per se, but rather that it is needed specifically for identities that continue to seem possibly false after they are believed.

7. Kripke Shows That Physicalists Don't Fully Believe Their Physicalism

In my view, Kripke's actual ad hominen argument poses a genuine problem for physicalism—a problem that isn't addressed by the physicalist response to the orthodox interpretation. In effect, Kripke shows that with mind-brain identities there's no gap between an appearance of possible falsity and a belief in actual falsity. If you believe a mind-brain identity to be true, then it can't appear possibly false to you, due to the absence of qualitative surrogates. So its appearing possibly false is incompatible with your believing it.

I think that Kripke is quite right about this. In sections 9 and 10 below I shall aim to show that there are no gaps in his line of reasoning. But first, in this section and the next, I want to clarify the larger dialectical situation.

I say that Kripke demonstrates that, if a mind-brain identity appears possibly false to you, then you can't believe it. And I accept that mind-brain

identities do appear possibly false to physicalists. However, I don't think that this disproves physicalism. For there is another option open to physicalists, apart from accepting that physicalism is false. Instead they can admit that they don't fully believe their physicalism. That is, they can agree with Kripke that mind-brain identities won't appear possibly false to someone who fully believes them—and so conclude, from the fact that such identities do strike them as possibly false, that they don't fully believe them.

This might seem tantamount to abandoning physicalism itself. But this need not follow. The idea would be to disbelieve physicalism at an *intuitive* level, while continuing to be committed to it at a *theoretical* level. There are plenty of good models for this kind of doxastic split personality. Consider the familiar Muller-Lyer illusion. At a theoretical level, we know that the lines are the same length. But at a more intuitive level of judgement they strike us as of different lengths. Nor is this kind of set-up restricted to cases involving perceptual illusion. At a theoretical level, I am entirely convinced that there is no moving present and the B-series description of reality is complete. But at an intuitive level I can't stop myself thinking that I am moving through time. At a theoretical level, I am persuaded that reality splits into independent branches whenever a quantum chance is actualized. But at an intuitive level I can't shake off the belief that there will be a fact of the matter about whether the Geiger counter will click in the next two seconds or not. And so on.

Physicalists can appeal to the same model in the mind-brain case. At a theoretical level, they fully accept mind-brain identities. But at an intuitive level something is stopping them embracing these beliefs.

This will then allows them to explain the appearance of possible falsity attaching to claims like 'pain = C-fibre stimulation', consistently with their theoretical commitment to these identities. This appearance of possible falsity arises at the intuitive level—at bottom it is simply the *intuitive* thought that pains and C-fibre stimulation are *actually* distinct. That's why it seems intuitively that there could be beings with C-fibre stimulation but not pain. At the theoretical level, on the other hand, possible falsity will be ruled out—once we keep it firmly in our theoretical minds that pain *is* C-fibre stimulation, we will find no room for the thought that 'they' might come apart, and so will dismiss the possibility of C-fibre stimulation without pain.

At both levels, then, Kripke's argument is respected. There's no question of simultaneously believing pain is C-fibre stimulation and in the same mode also thinking they mightn't be identical. We only think they might be different at the intuitive level where we think they *are* different; at the theoretical level where we think they are the same we don't think that they might be different. (Note that at neither level do we get an appearance of *contingency*—an appearance combining actual truth and possibly falsity. That I take to be ruled out by Kripke's argument. Rather we get an impression of actual falsity—a fortiori possible falsity—at the intuitive level, and an impression of actual truth—and so no possible falsity—at the theoretical level.)

I don't take this line of response to Kripke to weaken physicalism. Indeed I think that it is positively advantageous for physicalists to recognize that they are in the grip of a persistent dualist intuition. It is widely supposed that orthodox physicalism leaves us with some kind of 'explanatory gap' and so that some kind of extra resource—some new way of thinking about the physical world, perhaps—is needed fully to vindicate physicalism. The idea that we all are stuck with a persistent intuition of dualism casts a new light on this issue. Perhaps there is nothing more to the 'explanatory gap' than the strong intuition that C-fibre stimulation is one thing and pain another—that in itself would make us wonder *why* this extra feeling should be present whenever C-fibres are stimulated. If this is right, then physicalism won't need any extra resources to bridge the 'explanatory gap'. It will be enough to recognize that we are all in the grip of an intuition that physicalism is false.

Of course, this analysis does leave us with the question of *why* we should all be subject to this dualist intuition. A fully satisfactory physicalism would need to explain why physicalism is so hard to believe intuitively. Still, there seems no principled difficulty here. The current literature contains a number of suggested explanations for such an intuition of dualist distinctness. (Bloom 2004, Melnyk 2003, Papineau 1993, 2002, 2006.) However this is not the place to pursue this issue. My present concern is only to show *that* there is an intuition of distinctness. The further question of *why* this intuition should arise can be left for another occasion.

8. A Final Exegetical Point

I say that Kripke has a good argument to show that even physicalists intuitively disbelieve physicalism. Let me lay out the argument explicitly.

Argument A

(1) If you fully believe 'pain = C-fibre stimulation', then this can't appear possibly false to you.
(2) 'Pain = C-fibre stimulation' does appear possibly false, even to physicalists. So
(3) Even physicalists don't fully believe 'pain = C-fibre stimulation'

In the final two sections I am going to consider possible objections to this argument. But first I want to consider one final exegetical point. It might be objected that this can't possibly be Kripke's argument, because Kripke is arguing for the metaphysical conclusion that physicalism is false, whereas this arguments leads us only to the psychological conclusion that physicalists don't believe their physicalism.

This is a reasonable point. Even so, I would argue that my reconstruction fits nearly everything that Kripke says in the last pages of *Naming and Necessity*. After all, it would certainly make a kind of sense for Kripke's argument to have a psychological conclusion. For there is no question, I take it, but that his argument hinges crucially on a psychological *premise*, namely, that mind-brain identities 'appear contingent'. Remember, Kripke's argument isn't that these identities *are* contingent—he concedes from the start that this would beg the question against physicalism. Rather, he immediately retreats to the claim that they 'appear contingent', and starts arguing from there. Given this, it would scarcely be surprising that his argument should end up with a psychological conclusion. Indeed we might well wonder how it could end up with anything else. How could a psychological premise, that the reality *appears* a certain way to people, possibly imply anything about reality itself?

Still, I concede that Kripke does present his argument as an objection to physicalism, not just as a piece of psychological diagnosis. In this connection, it is interesting to ask how he manages to bring the psychological analysis to bear on the metaphysical issue. As far as I can see, this only happens in the last three sentences of the book. Until then he restricts himself to the ad hominem claim that 'identity theorists' have no good way of explaining the 'appearance of contingency' consistently with their physicalism. But in the final paragraph he adds the suggestion that this 'tells heavily against the usual forms of materialism. Materialism . . . must hold that . . . any mental facts are "ontologically dependent" on physical facts in the straightforward sense of following from them by necessity. No identity theorist seems to me to have made a convincing argument against the intuitive view that this is not the case.'

We can construe Kripke as here arguing along something like the following lines:

(i) intuitively it doesn't seem that the physical facts necessitate the mental facts;
(ii) unless this intuition can be explained away, it is evidence that physicalism is false;
(iii) this intuition cannot be explained away, so
(iv) (there is evidence that) physicalism is false.

If this is how Kripke reasons right at the end, then the rest of his anti-physicalist argument can be seen as a defence of the claim (iii) that the anti-physicalist intuition cannot be explained away on the 'qualitative surrogate' model—the appearance of contingency attaching to mind-brain identities is not an illusion akin to that attaching to 'heat = molecular motion'.

Of course, if this is how Kripke is arguing at the end, it is not a particularly strong form of argument. As he himself allows, the inapplicability of the 'qualitative surrogate' model doesn't necessarily mean that there is no other way of explaining away the anti-physicalist intuition.[4] For instance, consider the issue from the perspective developed in the last section. I there argued that the

earlier part of Kripke's argument already forces physicalists to recognize that the so-called 'appearance of contingency' is simply an intuition of actual falsity, and that a fully satisfactory physicalism therefore needs an explanation of why this intuition should arise even thought it is false. If physicalists succeed in finding such an explanation—and, as I said, a number of such explanations are on offer in the current literature—then they will automatically block Kripke's final metaphysical move, by showing that the anti-physicalist appearance of contingency can indeed be 'explained away'.

It might seem too easy to respond to Kripke's final move simply by agreeing that we intuitively disbelieve physicalism, and then seeking to explain why this should be so. But in truth this is an entirely reasonable move. At bottom, Kripke's final argumentative step is of this form: p seems false, so p is false. This is not a particularly strong form of argument. Many true things seem false. The natural response to any argument of this form is surely simply to explain why we should be inclined to disbelieve p even though it is true.

In effect, Kripke restricts himself to considering a different kind of explanation for the anti-physicalist intuition. He considers the hypothesis that claims of brain-mind necessitation strike us as counterintuitive because we muddle them up with other claims (their 'qualitative surrogates') that really aren't necessary—and he quite rightly points out that this hypothesis won't work in mind-brain cases. But it's scarcely as if the only—or even the most natural—way to explain a mistaken counterintuition about some truth is to show that we have muddled the truth up with some other claim that really is mistaken. After all, we are surely perfectly capable of directly disbelieving truths, even when we don't muddle them up with other claims.

9. Is Full Belief In Mind-Brain Identities Incompatible With An Appearance of Possible Falsity?

Let me now return to argument A. This may not culminate in the metaphysical anti-physicalist conclusion that Kripke was aiming at. But it is a striking argument for all that. It is certainly worthy our attention if Kripke's analysis shows that physicalists must intuitively disbelieve their physicalism. In the rest of this paper I want to show that this really does follow from the considerations that Kripke adduces.

I repeat Argument A for ease of reference.

(1) If you fully believe 'pain = C-fibre stimulation', then this can't appear possibly false to you.
(2) 'Pain = C-fibre stimulation' does appear possibly false, even to physicalists. So
(3) Even physicalists don't fully believe 'pain = C-fibre stimulation'.

In this section I shall focus on premise (1). Is full belief in mind-brain identities indeed incompatible with any appearance of their possible falsity? I shall consider four possible reasons for doubting this premise. Then in the next section I shall turn to some other responses to Argument A.

(i) *Does Everybody Know About The Necessity of Identity?*

Some might wish to query premise (1) on the grounds that it presupposes that ordinary thinkers are familiar with the necessity of identity. At bottom, the argument for premise (1) is that anyone who thinks that pains and C-fibres are actually identical will be in no doubt that they are necessarily so, given the absence of confusing qualitative surrogates in mind-brain cases. Still, what about people who don't know about the necessity of identity? After all, it was a great achievement of Kripke's to convince philosophers of the necessity of identity in the 1970s. So surely it would be easy for people who don't know of Kripke's work to believe fully that pain is C-fibre stimulation and yet not think that this is necessary?

I find this suggestion unpersuasive. After all, it is not as if Kripke changed the first-order modal thinking of ordinary people, as oppose to changing the explicit theories about modality articulated by philosophers. I would say that ordinary people have always thought that if Cicero is Tully, then this is necessarily so—except in cases where they muddle up this singular proposition up with some other claim. Similarly, I would say, any ordinary person who fully believes that pains are C-fibre stimulation will therewith believe that this is necessarily so—given that in this case there is no question of muddling this up with some other claim.

Furthermore, we can consider the case of more sophisticated physicalists who have read Kripke and do know about the necessity of identity. These thinkers at least will surely be disposed to move from full belief that pain is C-fibre stimulation to the conclusion that they can't possibly come apart. But, to run through the argument once more, it does appear even to these more sophisticated thinkers that pain can come apart from C-fibre stimulation; so they can't fully believe that pain is C-fibre stimulation; so something must be stopping them from believing their physicalism. And if something is stopping even sophisticated Kripkean physicalists from believing their physicalism, we can surely infer that the same barrier is also present in less sophisticated thinkers.

(ii) *Two Kinds of Imagination*

It is widely supposed that a physicalist answer to Kripke's challenge lies in the fact that we have two quite different ways of imagining the brain states that physicalists identify with conscious states.[5] We can imagine such brain states

perceptually, by imagining what it would be like to observe them (by imagining the neurographic readings produced by C-fibre stimulation, perhaps). But we can also imagine these states sympathetically, by imaginatively recreating the conscious states with which they are putatively identical (by imagining what it is like to be in pain, so to speak). Physicalists hold that what is being imagined in such cases is one and the same state, imagined either perceptually or sympathetically. But the two faculties of imagination are clearly themselves independent, in the sense that one can be exercised without the other. The suggestion is then that this independence of faculties itself enough to account for the 'appearance of contingency'. Because we can imagine the state perceptually without imagining it sympathetically, this creates the (illusory) impression that the brain state could exist without the conscious state.

I don't think that this succeeds in explaining the appearance of possible falsity. I accept, of course, that the two kinds of imagination mean that it an a posteriori matter that pain is C-fibre stimulation. Somebody might have the abilities to imagine C-fibre stimulation perceptually and to imagine pain sympathetically, but lack any evidence that the two states are the same. Moreover, this means that it is certainly *conceivable* that pain not be C-fibre stimulation, in that somebody could well suppose this without conceptual contradiction. Still, a central point I have aimed to establish in this article is that such conceivability does not by itself guarantee any appearance of contingency in people who already *believe* the relevant identity claim. You can recognise that a $\neq$ b is conceivable, and that other people can coherently believe this, but this needn't make you yourself think make the claim might have been false. Suppose you believe firmly that Cicero is Tully (and don't muddle this up with any other proposition). You'll still accept that there is nothing conceptually contradictory in the claim 'Cicero $\neq$ Tully' and that other people might indeed believe as much—but this won't create any impression in you yourself that Cicero mightn't have been Tully. (What are you supposed to think here? That Cicero might not have been himself?)

Does it make a difference that that in the mind-brain case we have perceptual and sympathetic imagination, rather than mere 'symbolic' imagination? Symbolically imagining Cicero without Tully is a far more anaemic exercise than perceptually imagining a brain state without sympathetically imagining the corresponding conscious state. Perhaps the more full-blooded imaginative exercise will create an appearance of contingency, even if the symbolic exercise does not.

But I don't see that the extra graphicness of perceptual and sympathetic imagination makes a difference. Suppose that I grow up thinking of 'Presley' visually, as the man with the pout and the long black hair, and of 'Elvis' aurally, as the man with the arresting voice. And then, later on, I discover that they are one and the same man. Now, I can still imagine someone with that pout and hair but without that voice. However, this act of imagination isn't going to make me feel that *Elvis* mightn't have been *Presley*, once I know that they are in fact the

same man. (What am I suppose to think here? That Elvis might not have been himself?)

Of course there is a real possibility here—a world where the person who sounds like that doesn't look like that. This may not be a world where Elvis isn't Presley, but it is clearly a genuine possibility for all that. So mightn't the impression of mind-brain contingency arise from an analogous source? Maybe this impression come from our awareness that the state with certain perceptual effects (producing those neurographic readings) might not appear sympathetically as it does (feeling like pain).

But this is where we came in, with Kripke's original argument. Mind-brain identities don't work like claims involving descriptions. Elvis might have cut his hair, but pain can't be pulled apart from its appearance. So when physicalists consider a scenario in which C-fibre stimulation does not appear sympathetically as it does, they aren't merely supposing that C-fibre stimulation (ie pain) might lack some contingent feature that it actually displays—rather they are supposing that C-fibre stimulation might appear without pain itself. And Kripke's original point was that this supposition can't explain how *physicalists* can think that pain and C-fibre stimulation might come apart, since it presupposes that C-fibre stimulation and pain are distinct states, contrary to their physicalist view.

(iii) *Descriptive Content on the Right-Hand Side*

Maybe pain can't be pulled apart from its appearance, but C-fibre stimulation arguably can. It is quite natural to think of this state, and of brain states generally, as physical kinds which only contingently have the characteristic causes and effects (including neurographic readings) by which we identify them. This is a familiar view of scientific kinds in general, associated with the Ramsey-Carnap thesis that theoretical terms are implicitly defined as referring to those entities that satisfy the assumptions of some surrounding theory.

This then suggests an alternative explanation for the appearance of possible falsity associated with claims like 'pain = C-fibre stimulation'. In this world the C-fibre role, so to speak, is realized by some physical kind X, and pain is identical to this kind X. But it is perfectly possible that the C-fibre role might have been be realized by some other kind Y, that is, by something other than pain. True, this isn't the possibility that *C-fibre stimulation* is not pain, if 'C-fibre stimulation' rigidly designates state X. Even so, perhaps an awareness of this possibility, that something other than pain might have realized the C-fibre role, could be responsible for the impression that 'pain = C-fibre stimulation' is contingent, even among those who firmly believe it.

I don't think that this works either. I am happy to agree that even those firmly committed to 'pain = C-fibre stimulation' can recognize the possibility that something other pain might have played the C-fibre role. But I don't think

that this is enough to explain the familiar impression that C-fibre stimulation might not have been pain.

When it strikes people—committed physicalists included—that C-fibre stimulation might not have been pain, their thought isn't just that some other (pain-free) state might have realized the C-fibre role. Surely they also think that the C-fibre role might have been realized just as it is, and yet pain might have been absent. That's surely the intuition that Kripke draws our attention to: everything in the brain could have been just as it is, all the way down, and yet pain might still have been absent.

Committed physicalists still seem to me to lack the resources to explain *this* intuition. It is incompatible with their belief that the actual brain facts fix the mental facts. True, they can suppose, compatibly with that belief, that the actual brain facts might have been different, and in particular that the C-fibre role might have been realized by some pain-free physical kind. But this supposition isn't going to help physicalists to think that the brain facts might have been just the same, and yet pain have been absent. Yet this is the intuition that Kripke challenges them to explain, and which still seems quite incompatible with their physicalism.

(iv) *Mightn't The Identity Have Turned Out to be False?*

Some readers might still be wondering whether there isn't a simple explanation for the appearance of contingency. Even after you come to believe that pain = C-fibre stimulation, won't the mere fact that this is an a posteriori truth mean that you will still be aware that it *might have turned out otherwise*? Mightn't the evidence have shown that pain wasn't C-fibre stimulation after all, but some other brain state? And won't this in itself account for your persistent feeling that C-fibre stimulation mightn't have been pain, even though it actually is?

No. As I have been stressing for some while, a posteriority by itself need not create an impression of contingency, at least not in those who fully believe some necessary claim. Suppose you are now certain that Cicero is Tully (and moreover aren't inclined to muddle this claim up with some other proposition). Will you think that Cicero (that very person) might not have turned out to be Tully (that very person)? Surely not. How could Cicero have failed to be Tully? True, you might hold that the evidence showing that Cicero is Tully might have failed to come to light, and that people would therefore have remained ignorant of this identity. But that's the possibility that people might not have known that Cicero is Tully, not the possibility that 'they' aren't the same person.

I say that, if you firmly believe some necessary truth, and there are no qualitative surrogates around, then you won't think that this truth might have turned out otherwise. You can think that people might have failed to discover this necessary truth, but that is different. So if you firmly believe that pain is C-fibre stimulation—that they are one and the same state—then you will have no room left for the thought that they might have turned out to be different.

Once more the persistence appearance of possible falsity in physicalist believers resists explanation.

10. Is The Intuition of Possible Falsity Simply Due to Ignorance?

Maybe the reason 'pain = C-fibre stimulation' appears possibly false to physicalists is not that they think it might have turned out differently, but rather that they don't yet know how it will turn out.

After all, nobody really believes the specific claim that pain is C-fibre stimulation. It is well-known that this is not a good account of the complex psycho-physiological data on pain. And the same could be said of most other attempts to equate specific conscious states with specific brain states. At best these equations offer hypotheses for further investigation. That's why philosophers typically invoke an example of a mind-brain identity—pain = C-fibre stimulation—that they know not to be adequately supported by the evidence. For in fact there aren't any good examples of well-established mind-brain identities to use instead.

Accordingly, most contemporary philosophical physicalists restrict themselves to a generalized physicalism. They admit that brain science is still in its infancy, and that we are not yet in a position to assert any claims of the form 'M = P' where 'M' names some specific mental state and 'P' some definite brain state. At best we can assert that for each mental state there is *some* physical state to which it is identical, but which we don't yet know about.[6]

This then suggests a different explanation for the impression of possible falsity that attaches to any specific mind-brain identity of the form 'M = P'. Maybe any such claim strikes us as possibly false simply because we don't believe it, and so of course attach some positive credence its being false.

If this is the right diagnosis, it suggests that Argument A is perfectly sound, but quite unsurprising. Let us agree that (1) if you fully believe 'pain = C-fibre stimulation', then this can't appear possibly false to you, and that (2) 'pain = C-fibre stimulation' does appear possibly false, even to physicalists. From this it certainly follows (3) that even physicalists don't fully believe 'pain = C-fibre stimulation'. However, on the current suggestion, this isn't because some deep cognitive obstacle is preventing them from embracing physicalist claims that are supported by overwhelming evidence. Rather, it's simply because there isn't yet any good evidence for any specific such claims. That's the only reason why physicalists don't believe them, and so of course think them possibly false. Or so at least the current suggestion goes.

However, I don't think that this suggestion amounts to an adequate response to argument A. It is perfectly true that specific claims of the form 'M = P' are currently under-evidenced, and that as a result few physicalists actually believe any specific such claims. But even so I think that reflection on Argument A gives us good reason to suppose that anti-physicalist intuition goes deeper than this temporary ignorance.

For a start, note that the suggested response implies that, once physicalists *do* have good evidence for specific claims of the form 'M = P', such claims will cease to strike them as possibly false. I'm not persuaded of this. Imagine that physicalist scientists have now uncovered overwhelming evidence that in the actual world pain always goes hand in hand with some brain process K. Wouldn't it still strike them that there could be a possible being with process K yet no pain? I would have thought that this intuition would remain. But then Argument A shows once more that something would be stopping these scientists fully believing the physicalist claim 'pain = process K', in the face of evidence that we would expect to persuade them of this—for, if they were fully to believe the identity, then it wouldn't still appear possibly false to them.

Effectively the same point can be made without any counterfactual assumptions about how things would strike phsyicalists who had more evidence. Instead of focusing on specific mind-brain identities 'M = P', consider instead the generalized physicalist belief referred to above, that for each mental state there is *some* physical state to which it is identical. Most contemporary physicalists take themselves *already* to have conclusive evidence for this generalized claim, even if not for more specific identities (cf Papineau 2002 ch 2). But if such physicalists fully believe this generalized physicalism, then the Kripkean considerations imply that ought to have no remaining room for any thought that this generalized physicalism could be false. But they do.

After all, most contemporary physicalists will readily admit that zombies do at least appear possible, even if in reality they are not. It certainly seems at first pass that a being could share all our physical properties and yet lack any conscious life. But this appearance itself is incompatible with a commitment to a generalized physicalism. If you believe that each of our mental properties is identical to one of our physical properties, then surely you must think that a being with *all* our physical properties must have all our mental properties too. So what are you doing thinking that such a being might lack our mental properties? (If you fully believe that your son is one of the children in that group, even though you can't see which, do you think that all those children could go into the classroom but your son not be there?)

So now we have a generalized version of Argument A.

1. If you fully believe that each mental state is identical to *some* physical state, then zombies ought not to appear possible to you.
2. Zombies do appear possible, even to physicalists. So
3. Even physicalists don't fully believe their generalized physicalism.

I conclude, once more, that physicalists are in the grip of an intuition of dualism that prevents them from fully believing their generalized physicalism.

Let me conclude with a concession. Maybe the appearance of falsity attaching to physicalist doctrines will indeed fade away one day. In a future where physicalist views have become part of educated common sense, then perhaps

it will no longer seem obvious that brain facts cannot necessitate conscious facts. Thus zombies may cease to appear possible. (Look—they've got all my mind/brain states—how can they not be conscious?) And perhaps even specific mind-brain identities may cease to appear possibly false. (How could there be process K without pain?—that's what pain *is*.)[7]

I have been arguing that the Kripkean considerations show that there is some barrier to physicalist belief—something is stopping even those who profess physicalism from fully embracing their own doctrines. But I haven't said anything about the nature of this barrier to belief. Maybe it lies in some deep architectural feature of the mind, and so will indefinitely remain a source of cognitive illusion (cf the Muller-Lyer lines). On the other hand, maybe it is a relatively shallow phenomenon, due to nothing more than the unfamiliarity of genuine physicalism, and so will disappear once this view ceases to seem so strange. To decide between these options, we would need to know more about the mechanism behind the resistance to physicalism, and that is an issue I have avoided in this paper.

Still, I hope I have done enough to show that there is as least some *current* resistance to physicalist doctrines, even among those who sincerely profess physicalism. Contemporary physicalists may not be fully committed to any specific mind-brain identities, but they do at least profess to believe in a generalized physicalism. Yet zombies still strike them as intuitively possible, even if this intuition is fated to fade away in the future. I say that they wouldn't currently have this zombie intuition, if they fully believed their generalized physicalism. So something is currently stopping them from fully believing this.[8]

Notes

1. See for example Chalmers (2006). It should be noted that Chalmers and some other advocates of two-dimensional semantics take 'unstable primary intension' to be a more general notion than that of a contingent descriptive content and would object on these grounds that my formulation of principle (I) coarsens their interpretation of Kripke. However, I don't think that this point matters to any of the arguments to follow—and it is much easier to explain things in terms of 'descriptive contents' than 'unstable primary intensions'.
2. I am here rehearsing the 'a posteriori' physicalist line. Different a posteriori physicalists have different views about the extent to which principle (I) fails, but they all agree that mind-brain identities in particular constitute exceptions. (Loar 1997, Hill 1997, Levine 2001, Papineau 2002.)
3. A couple of years ago I offered my colleague Keith Hossack the standard physicalist response to the two-dimensional reading of Kripke. He responded 'That might explain an appearance of possible falsity, but it doesn't explain the appearance of *contingency*'. It was this remark that got me thinking about the proper way to interpret Kripke.
4. 'That the usual moves and analogies are not available to solve the problems of the identity theorist is, of course, no proof that no moves are available' p 155.
5. See Nagel 1974, footnote 11, Hill 1997, Hill and McLaughlin 1999.

6. True, many contemporary physicalists believe something yet weaker, since they formulate physicalism in terms of supervenience rather than identity. But let me continue to focus on identity, in line with Kripke's original discussion of physicalism. This is for expository convenience only—all the points made in this paper apply equally to supervenience versions of physicalism.
7. Stephen Yablo suggests that future discoveries may well undermine anti-physicalist intuitions: 'Am I the only one who feels the intuition of zombies to be vulnerable in this way?' (op cit, p 119).
8. A version of this paper was given at a Pacific APA symposium in April 2007. I would like to thank Tyler Doggett for his very helpful comments. Earlier versions were delivered to seminars at Birmingham, Cambridge, North Carolina at Chapel Hill, Sussex, Berlin, Nottingham and King's College London. I would like to thank all those who responded on those occasions, especially David Chalmers, Keith Hossack, Christian Nimtz, Mark Sainsbury, and Gabriel Segal.

References

Bloom, P. 2004. *Descartes' Baby*. New York: Basic Books.

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

Chalmers, D. 2006. 'The Foundations of Two-Dimensional Semantics,' in García-Carpintero, M. and Macià, J. (2006).

García-Carpintero, M. and Macià, J. 2006. *Two-Dimensional Semantics*. Oxford: Oxford University Press.

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

Hill, C. and McLaughlin, B. 1999. 'There are Fewer Things in Reality Than Are Dreamt of in Chalmers' Philosophy,' *Philosophy and Phenomenological Research* 59.

Kripke, S. 1980. *Naming and Necessity*. Oxford: Blackwell.

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

Loar, B. 1997. 'Phenomenal States: Second Version,' in Block, N., Flanagan, O., and Guzeldere, G. (eds), *Consciousness*, Cambridge, Mass: MIT Press.

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.

Nagel, T. 1974. 'What is it Like to be a Bat?' *Philosophical Review* 83.

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

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.

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