The Significance of Intuitions of Contingency for the Mind-Body Problem

by

Judith M. Feldmann

B.A. English
Franklin & Marshall College, 1990

M.A. Philosophy
Tufts University, 1992

Submitted to the Department of [Linguistics and] Philosophy
in Partial Fulfillment of the Requirement for the Degree of

Doctor of Philosophy
with the thesis in the field of Philosophy

at the
Massachusetts Institute of Technology

February 1997

Signature of Author:.....
Department of Philosophy
October 10, 1996

Certified by: .....
Robert Stalnaker
Professor of Philosophy
Thesis Supervisor

Accepted by: .....
Alexander Byrne
Chairman, Department Committee on Graduate Students

The Significance of Intuitions of Contingency for the Mind-Body Problem

by

Judith M. Feldmann

Submitted to the Department of Philosophy
on October 10, 1996 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy
with the thesis in the field of Philosophy

ABSTRACT

I discuss the significance of intuitions of contingency for the Mind-Body problem. The intuitions are of the alleged contingency of the relation between the mind and the brain. Chapter one deals with White's property dualism argument, in which he argues that we are committed to property dualism if there is no a priori identification of the mind and brain. I argue that his argument is unsound given the falsity of a semantic principle, the effect of which is to conflate our having no a priori identification with our having successful intuitions of contingency. I also argue that Loar makes the same conflation in his reply to similar arguments.

Chapter two deals with the intuitions as they arise in arguments for the explanatory gap, an allegedly "merely epistemic" problem for physicalism. Proponents of the gap argue that since intuitions are a matter of what is conceivable, they have no metaphysical implications; rather they show only what is epistemically possible. I argue that this reply involves a flawed view of conceivability. The intuitions have implications (whether these are true or false). The real issue lies in whether what is claimed to be conceivable is genuinely conceivable. I conclude that the intuitions of contingency pose no merely epistemic problem over and above the original metaphysical challenge.

In chapter three, I propose a compositional view of the mind as an empirical thesis that would account for one sort of intuitions of contingency. First, I classify the intuitions into five types. The first three types support either substance or property dualism; the fourth supports functionalism; and a fifth supports the proposed view, according to which a mental property such as pain is the property of being composed of physical matter that instantiates certain physical properties of the brain. I contrast the view these intuitions support with Searle's biological naturalism. Since Searle claims to rely on one of the first three sorts of intuitions, he cannot go on to claim as he does that biological naturalism avoids dualism. I argue that my proposed view should be welcomed, over Searle's view, by materialists.

Thesis Supervisor: Robert Stalnaker
Title: Professor of Philosophy

Chapter 1

The Property Dualism Argument:

Intuitions of Contingency and A Posteriori Identifications

0. Introduction

The property dualism argument, discussed recently by Stephen White, is meant to show that certain a posteriori claims that identify mental and physical entities entail the truth of property dualism. This argument relies on a powerful semantic principle, that if there are no true a priori statements identifying mental and physical entities, there must be at least two properties of the entity in question that serve as distinct modes of presentation of that entity. I will argue that this principle is false, and so the property dualism argument is unsound. I will also argue that the argument for this principle fails in its inference from there being no true a priori identity statement to its being conceivable that the a posteriori identity statement is false, an inference which makes some sense when the identification involves expressions that are non-rigid, but which does not make sense when the identification involves rigid designators. This slide may serve to contribute to the apparent plausibility of the principle by conflating it with a weaker principle: if we can coherently conceive of the falsity of an a posteriori identity statement, then distinct properties serve as modes of presentation of the object(s) in question. This slide is, moreover, arguably the source of the problematic view that what is conceivable may nevertheless be impossible, a view which is offered by Loar in partial response to the property dualism argument.

First, I will present the property dualism argument (hereafter, PDA) as it is formulated by White. Then, I will offer a revised interpretation of the PDA, in order to render the argument valid in a way which does not beg the question against physicalism. I will then go on to examine failed criticisms of the PDA. Among these unsuccessful criticisms will be Brian Loar's criticism that the general principle of the PDA rules out direct reference; while I agree with Loar that the problem for the PDA involves ruling out direct reference, I disagree with his claim that admitting this involves admitting that what is conceivable is nevertheless impossible. He commits

himself to this because he inherits the slide mentioned above, from there being no true a priori identity statement to its being conceivable that the a posteriori identity statement is false. Finally, I will explain in more detail what I think is the real problem with the argument.

1. The Property Dualism Argument

1.1 The Property Dualism Argument as formulated by White

In 'Curse of the Qualia', White presents the PDA as a means to defending a priori functionalism against an attempt to accommodate it to intuitions about the possibility of absent qualia. According to these intuitions, it seems metaphysically possible that the mental and physical are distinct: one could instantiate the relevant physical properties, and yet instantiate no (phenomenal) mental properties whatsoever. White phrases his argument in terms of the a posteriori nature of the identification of the mental with the physical: the identification, if it is correct, is one we would need to discover. It may seem, then, that he need not rely on a discussion of intuitions of contingency. However, as we shall see below, in his argument for a crucial premise, White claims to conclude from there being no true a priori identity statement to its being conceivable that the a posteriori identity statement is false. There are deep problems with making such an inference, as we shall see; but on its face it may seem plausible. For, it may seem that most statements we need to discover are contingent statements, which are conceivably false. However, the issue gets more complicated once we admit (with Kripke) that we can discover a posteriori necessary truths. Nevertheless, it may be that there will be a feeling of contingency to whatever we discover a posteriori. However, some of these intuitions of contingency, those that are associated with necessary a posteriori truths, will need to be explained away, if we are to salvage the truth of the identity statement.[1]

What the PDA does, then, is to conjoin the fact that there is no a priori identification with *successful* intuitions of contingency, the effect of which is to

guarantee the content of the intuitions as correct, i.e. to imply the truth of some form of dualism. White's ultimate goal is to show that functionalism combined with some sort of physicalism in order to account for the phenomenal nature of qualia falls prey to the PDA, and so, barring a willingness to commit to property dualism, presupposes the view of a priori functionalism upon which it is meant to improve. In what follows we won't be concerned with his use of the PDA to support a priori functionalism, but rather with the PDA itself, i.e. the argument that if there are no true a priori statements identifying mental and physical entities, we are committed to property dualism.

How does White argue for the above conditional? He makes the following remarks, on pp. 92-94:

> The general principle is that if two expressions refer to the same object and this fact cannot be established a priori, they do so in virtue of different routes to the referent provided by different modes of presentation of that referent. ... [T]he natural candidates for these modes of presentation are properties. ... Since there is no physicalistic description that one could plausibly suppose to be coreferential a priori with an expression like 'Smith's pain at t', no physical property of a pain (i.e., a brain state of type X) could provide the route by which it was picked out by such an expression . ... If there are no topic-neutral expressions [i.e. expressions that are neither mentalistic nor physicalistic] that are at least coreferential a priori with such mentalistic descriptions as 'Smith's pain at t', then these mentalistic descriptions refer in virtue of a property distinct from that in virtue of which any physicalistic or topic-neutral expression refers. Such a property or feature could only be regarded as an irreducibly mental entity.

From the above remarks, the PDA seems to be the following[2]:

Assumption: If an expression is not mentalistic, it is physicalistic.

1. Suppose there are no true a priori statements identifying the mental with the physical, i.e. suppose that for any mentalistic expression M (e.g., 'Smith's pain at t') and for any physicalistic expression P (e.g., 'Smith's c-fiber stimulation at t'), if the claim <M = P>[3] is true, it is true only a posteriori. [Giving rise to intuitions of contingency]

2. If two expressions, a and b, refer to the same entity and the sentence <a=b> cannot be known to be true a priori, the expressions refer in virtue of different senses, i.e. modes of presentation of the referent, where these are construed as properties of the referent. [White's general principle]

(from 1 and 2) 3. There is a property, P1, the sense of M; and P1 is not the sense of any physicalistic expression.

4. If M is coreferential a priori with only mentalistic expressions, the sense of M is an irreducibly mental property.

(from 3 and 4) 5. P1, the sense of M, is an irreducibly mental property.

Therefore: if there are no true a priori statements identifying mental and physical entities, we are committed to property dualism.

The argument as it stands, however, is valid only given its unsupported premise 4. Further, the content of this unsupported premise is so close to that of the conclusion, it is in danger of being deemed question-begging. We are asked to accept that for an expression to be coreferential a priori with only mentalistic expressions is for it to refer via an irreducibly mental mode of presentation. Why not, we should ask, affirm instead a weaker claim, that for an expression to be coreferential a priori with only mentalistic expressions is for it to refer via a mental property, leaving open the possibility that this property is also physical? (In particular, it may be also a physical property if we discover a posteriori that the mentalistic expression is coreferential with a physicalistic expression.) What, in particular, does White mean by 'mentalistic' and 'physicalistic' expressions?

White seems to make two assumptions concerning these expressions: first, that we associate some sort of description with each expression (perhaps just the expression itself), from which we can ascertain which property is the mode of presentation of the referent. Not only do we ascertain the relevant property in this way, but we are also able to ascertain what sort of property it is, i.e. whether it is mental or physical. And secondly, he must assume that physical and mental properties are mutually exclusive (or, as in premise 4, that mentalistic expressions refer to irreducibly mental properties). It is this second assumption which allows us to close off possibilities that should be left open.

That this assumption comes dangerously near to begging the question is, I think, clear. Take a paradigmatic mentalistic expression like 'Smith's pain at t', and suppose its mode of presentation is a feeling of painfulness. To get the conclusion that this feeling of painfulness is irreducibly mental, either White must assume that the modes of presentation of mentalistic expressions just are irreducibly mental properties, or he must assume that in particular, the feeling of painfulness is just obviously an irreducibly mental property. Either way, he begs the question against the opposing side. For what is at issue is not merely whether what is referred to by mentalistic expressions is identical to what is referred to by physicalistic expressions, but further whether any properties that are involved in ensuring this referential relation holds are themselves able to be accounted for by some physicalistic theory. The question is one of the status of mental and physical entities, whether these are objects or properties.

1.2 The PDA revised

I take it, then, that White must not presuppose that the descriptions we may associate with the senses of these terms must refer in virtue of one sort of property rather than another. If this is the case, however, he needs some substantive reason why mentalistic expressions can refer only in virtue of irreducibly mental entities: he needs further support for premise 4. Is there anything that could be supplied here, to render the argument valid in a manner that does not beg the question? I think something could be, namely a claim to the effect that if premise 4 is false, an infinite regress of properties emerges. The revised argument then becomes the following:

1. Where M is a mentalistic expression (e.g., 'Smith's pain at t') and P is a physicalistic expression (e.g., 'Smith's c-fiber stimulation at t'), if the claim $\langle M = P \rangle$ is true, it is knowable only a posteriori; i.e., M is coreferential a priori only with mentalistic expressions.

2. If two expressions, a and b, refer to the same entity and the sentence $\langle a=b \rangle$ cannot be known to be true a priori, the expressions refer in virtue of different senses, i.e. modes of presentation of the referent, where these are construed as properties of the referent.

3*. Any physical property has a physicalistic expression that refers to it, and any mental property has a mentalistic expression that refers to it.

Assume for reductio: Mental properties, i.e. the senses of mentalistic expressions, are physical properties.

(by assumption) 4*. The sense of M is some physical property, P1.

(from 3* and 4*) 5*. P1 has both a mentalistic expression, M1, (e.g. 'the sense of 'Smith's pain at t'') and a physicalistic expression, Q, referring to it.

(from 1 and 5*) 6*. The statement <M1 is Q> is knowable only a posteriori.

(from 2 and 6*) 7*. So M1 and Q have distinct senses.

(by assumption) 8*. The sense of M1 is a physical property, P2.

(from 3* and 8*) 9*. P2 has both a mentalistic expression, M3 (e.g 'the sense of 'the sense of 'Smith's pain at t''), and a physicalistic expression, Q1, referring to it.

(from 1 and 9*) 10*. The statement <M3 is Q1> is knowable only a posteriori.

(from 2 and 10*) 11*. So the expressions M3 and Q1 have distinct senses; and the argument can be seen to introduce properties ad infinitum.

12*. Such an infinite regress of properties is impossible.

Therefore, the assumption is false: No sense of a mentalistic expression is a physicai property.

This argument, I propose, is an improvement over the earlier version, in that it supplies a reason for thinking that the sense of a mentalistic expression cannot be a physical property. The question-begging premise 4 has been replaced with the weaker 3*, which seems on its face more plausible, and the content of 4 which lacked support has been supported by a reductio. Still, the argument may not seem convincing. I think the real problem with argument is the general principle, touted in premise 2. However, first I will discuss some other criticisms and contest that they ultimately do not succeed.

2 Failed criticisms of the PDA

2.1 Embracing the regress

One response to the above revised PDA is that we can remain physicalists by simply embracing the regress, and denying that it is impossible. If, given our views on the nature of language, we accept the general principle, then we will think that certain a posteriori identities depend on there being distinct modes of presentation of the identified referent; but such a view can remain physicalistic if each of these is in turn a physical property. Why should we suppose that such a regress is impossible? Perhaps we are shunning this possibility out of something akin to ignorant prejudice.

This response demands that the proponent of the (revised) PDA explain what is problematic about the regress; and there is something to be said for this skepticism. Is it transparently incoherent that there be such a regress of properties? Suppose one tried to argue that the regress of properties was incoherent because it involves making an incoherent claim about our understanding: to grasp the sense of a term, one needs to grasp an infinite regress of senses all at once. If this were the case, understanding would never get off the ground. However, it is unclear that to grasp a sense of a term, it is necessary to grasp simultaneously the sense of that sense. In other words, perhaps to understand the above claim that <M=P>, one need only grasp the senses of the expressions, M and P; he need not grasp the senses of those senses (though he may be required to grasp them in order to understand other claims). So this reason for finding the regress repulsive is not compelling.

However, while it is true that a regress of properties may not be transparently incoherent, still, I think it is at least implausible, if we look more closely at how it is meant to arise. I think there is something intuitively problematic and implausible (if not impossible) about the regress, which may push us to seek to undermine the argument elsewhere. What is problematic about such a regress is that it arises as a result of the general principle, which is a semantic principle. Perhaps there are infinite regresses of properties; but why should we think that they arise as a result of

our semantic connection to the world? Why not? one may ask. Why should a source such a this seem more insidious than any other? Could this be simply more blind prejudice? Admittedly, some will find the prospect of this sort of regress easier to countenance than others; but note that embracing the regress means embracing the fact that metaphysical possibilities grow at an astounding rate merely given our use — in fact, our misuse — of language. Since our semantic connection can and does change and evolve, this must mean that such regresses are propagated rather alarmingly, with almost each discovery. According to the proponent of the PDA, since we did not see a priori that an identity statement was in fact true, there are infinitely many properties. Are we to take it that had we used different expressions, different properties would have existed? Or that perhaps an infinite number of such regresses exist, and we happen to have latched onto one with the expressions we have chosen? Either route seems to multiply the metaphysical possibilities. For this reason, I do not think that embracing the regress is the best way for the physicalist to respond to the revised PDA. Clearly, in giving this reply I have not refuted those who are happy to embrace the regress. At any rate, whether or not we are tolerant of regresses, there are other problems with the argument, which seem to me to be prior to the regress question. The prior problem is whether we should accept the powerful general principle. Brian Loar is another who recognizes that this principle is problematic.

2.2 Brian Loar's rejection of the general principle

2.21 Loar's reply: phenomenal concepts

My reply to the above serves to point us to another deeper objection to the PDA, namely, why should we accept the general principle, its first premise? The regress arises as a result of our adherence to this principle, and if we find the regress at all problematic we should examine this source more closely. Brian Loar, in his paper 'Phenomenal States', argues that this principle is false because it rules out the possibility that some terms, e.g. mentalistic ones, refer directly, rather than in virtue of a higher-

order reference-fixing property. He offers a specific counterexample to the principle, to show how a term like 'pain', and moreover how our concept of pain, could refer directly to a brain state of, e.g., c-fiber stimulation. In arguing this way, he wants to claim that we can preserve the anti-physicalist intuition that mental entities are distinct from physical entities, and yet at the same time deny that this intuition has any metaphysical implications, i.e. deny that such a distinction in our concepts implies that there is any distinction in properties. It is a recognized consequence of his argument that conceivability is not a good test for possibility (Loar, p. 19); that this is a consequence will become clearer as we see the outlines of his view.

It will helpful, first, to lay out how Loar presents the property dualism argument. The first premise, according to Loar, is the anti-physicalist intuition: 'we can coherently conceive any given phenomenal quality in the absence of any physical-functional property and conversely' (Loar, p. 2). A few pages later, however, he makes a crucial change to the first premise. This first premise then becomes the following: 'Knowing that p, if p is conceived in physical-functional terms, never a priori suffices for knowing that q, if q is conceived of in phenomenal terms' (Loar, p. 4). This, he claims, is the 'anti-physicalist's basic intuition'; but note that this intuition is a more cautious one than the intuition that we can presently conceive of the falsity of the a posteriori identity statement. This distinction, and its ramifications, will become much clearer below. The second premise is White's general principle. Loar calls this 'the semantic minor premise': 'the cognitive independence of *conceptions* couched in phenomenal and in physical-functional terms [i.e. the aforementioned conceivability] implies the distinctness of the phenomenal and physical-functional *properties* that those conceptions connote or stand for' (Loar, p. 2). From these two premises, the conclusion follows that there are phenomenal properties that are distinct from the physical-functional properties, connoted by the terms in the original identity statement. (Loar at this point does not consider the possibility that these phenomenal properties could

themselves be identified a posteriori with other physical-functional properties, thereby giving rise to a regress, but here I shall assume for simplicity that he would not be willing to accept the regress.) This is Loar's initial characterization of the argument. His strategy is to accept premise 1, and deny premise 2.

Loar claims that it is easy to describe in outline what would falsify the second premise. In Loar's words, the form of his counterexample to White's general principle is as follows:

> A pair of concepts...may...converge on a property, may have that property as their common reference, in the following way. A recognitional concept can involve the ability to class together, to discriminate, things that have a given objective property. Say that if a recognitional concept is related thus to a property, the property triggers applications of the concept. Then the property that triggers the concept is the semantic value of the property, unmediated by a higher-order reference-fixer. Now suppose we have an independent account of what property a given theoretical concept refers to. Nothing prevents that property from being the property that triggers a given recognitional concept, and so the two concepts can converge in their reference despite their cognitive independence, the latter being a sort of brute psychological fact. (p. 7)

Loar's strategy is to present us with some means of referring which allows us to have an a posteriori identity without any leftover descriptional content which we must countenance as defining a distinct property. On his view, phenomenal concepts are concepts which recognize phenomenal properties, by demonstratively pointing to those properties. He claims that these phenomenal concepts, as recognitional concepts, refer directly to whatever state that triggers them. Rather than supposing that there is any descriptional content to our concept of pain, he claims that it refers directly to the underlying brain state. A phenomenal state, then, on this view, will be whatever state it is that triggers the relevant phenomenal concept; and so whatever it is that triggers our phenomenal concepts is alleged to be that feature of the state that we point to as the experiential way the state appears to us, i.e. the way it feels. According to Loar's picture, then, phenomenal concepts are involuntary in the sense that whenever we are

in a particular brain state, we are struck with the phenomenal concept which refers to that brain state in virtue of the triggering mechanism. Presumably, for Loar, the phenomenal concept itself must be some physical-functional state of the brain. Moreover, these phenomenal concepts are constituents in representational mental sentences; we employ them in genuine propositional thoughts about phenomenal states (Loar, p. 11). Loar wants to grant that knowing how a phenomenal state feels is knowing that it feels a particular way, rather than merely knowing how to categorize or classify certain sorts of states.

Furthermore, Loar argues that he can explain how it is that the antiphysicalist can remain convinced that the cognitive independence of his concepts of the mental and the physical implies that there are distinct properties: given the association of phenomenal concepts with imaging, phenomenal concepts seem more akin to phenomenal states themselves than do theoretical concepts (Loar, p. 10). This striking difference between the two sorts of concepts seduces us into thinking that they are actually concepts of different things, and that, say, purely physical-functional states need not have any phenomenal feel to them at all. The physicalist, however, while granting that our concepts of the two are cognitively independent, can deny the metaphysical implication, given that he sees how the referents of the radically distinct concepts converge; what is conceivable is nevertheless impossible. This, then, is how Loar accepts the consequence that conceivability does not imply metaphysical possibility: one can have the intuitions that the mental and physical must be distinct, and this distinction is genuinely imaginable, given the brute triggering of the phenomenal concept. Nevertheless, given that there is no descriptional content to our phenomenal concepts, we can draw no metaphysical conclusions from them; we must wait and see what it is that has triggered them in us, to learn of their true semantic value. Loar can admit that our theoretical concepts have metaphysical implications; but

our phenomenal concepts do not, since they merely serve to point blindly at whatever triggers them.

These, then, are the bare bones of Loar's view. First, I think we should grant that this does, on its face, provide a counterexample to the general principle, regardless of whether it captures the way our concepts of phenomenal states actually work. Loar's example shows how we may have two concepts which converge on a single referent without having distinct reference-fixing properties that provide modes of presentation of the referent. So he seems to have shown how cognitive independence, construed as arising from an a posteriori identity statement, can fail to imply a distinction in properties; and this is to grant the antecedent of White's principle and deny its consequent.

For suppose one tried to object that Loar does not falsify the general principle, but merely trivializes it. After all, Loar seems to grant there are nevertheless distinct properties of the referent: given the phenomenal concept, there is the property of being related to the referent by the triggering mechanism; and given the theoretical concept of c-fiber stimulation, there is some corresponding property of being the phenomenon that fits the appropriate description. But these two properties are distinct, so the objection goes, and one might worry that Loar has not falsified the principle. After all, the distinct properties of the referent still seem to arise as a result of the a posteriori identification.

Such an objection seems to stretch the notion of sense: for if the triggering mechanism is simply built into the notion of sense, so that any two terms that are triggered by separate mechanisms have distinct senses, then any two terms will have senses in this sense and so will refer in virtue of distinct properties. But then these properties will not obviously be the problematic mental properties that the PDA relies on. After all, this introduction of distinct properties will be one that applies wherever there are distinct names for an object, and as such will not be any new problem for

philosophers of mind. If one wants to argue, then, that Loar has only managed to trivialize the general principle, he must also admit that Loar has effectively rendered it useless. If we want to object to Loar's view we will have to find some other way of doing so.

2.22 An incoherent account of conceivability

While it is true that Loar falsifies the general principle by pointing out that it rules out a possible situation, he also takes it that his reply has the following consequence: what is conceivable is nevertheless impossible. In this section, I will argue that this consequence arises given Loar's conflation between our having intuitions of genuine contingency on the one hand, and there being no true a priori identity statement on the other. Moreover, it is understandable that he should conflate these, since White's argument for the general principle, as we shall see below, relies on just such a conflation. So it is not as though Loar is misinterpreting the general principle. Such a consequence, moreover, should be considered unwelcome. For what does it mean to claim that one can conceive of some impossible situation? While it may be true somehow that one can unwittingly believe contradictions, it is difficult to see how it could be true that one could *coherently* conceive of an impossibility.

However, Loar seems to want to claim that this is somehow possible. As he says, 'The need for a further premise [i.e. the general principle] naturally suggests that a physicalist can agree with the basic anti-physicalist intuition, i.e. that we can coherently conceive any given phenomenal quality in the absence of any physical-functional property and conversely. This is simply to accept that phenomenal representations are cognitively independent of physical and functional descriptions' (Loar, p. 2). What he then claims, as we saw above, is that our concepts that refer directly have no metaphysical consequences, just as, presumably, names that refer directly have no descriptional baggage. There is no meaning attached to a phenomenal concept, over and above its semantic value. This allows the anti-physicalist to

coherently conceive that phenomenal properties are distinct from physical-functional properties, even though this latter claim is necessarily false. But what does this mean? How can anyone coherently conceive of what is necessarily false?

Loar himself notes this objection, and dismisses it (though interestingly, there is no explicit recognition of this problem in his original published version of the paper):

> The present account implies that an apparent state of affairs may be conceivable even though no metaphysical possibility corresponds to it. For, on this account, we may coherently conceive that a physical-functional property **P** can be exemplified without phenomenal quality **Q** (as conceived in phenomenal terms), even though it is metaphysically necessary that **P** and **Q** go together, being identical. ... But it may be objected that, if conceivability is not a criterion of metaphysical possibility, our epistemic access to metaphysical modality is seriously in jeopardy. For what apart from conceivability gives us access? The objection apparently asks us to accept that we should regard conceivability as implying possibility even though we can give a straightforward account of certain concepts and their reference-determination that implies that conceivability does not in this case imply possibility. We are asked to accept this *even though* there are no objections to that account on psychological or reference-theoretic grounds. But the claim that the conceivability of certain things implies its possibility is surely undercut if we can show *how* we could coherently conceive it even though that conception failed to capture a possible state of affairs. (Loar, pp. 19-20)

I think that it is a serious objection to a view about our concepts that the view implies that our concepts have no metaphysical implications, in the sense that it implies that given our concepts, we can coherently conceive of impossibilities. That this is not a psychological or reference-theoretic objection fails to lessen its importance. Furthermore, it is unclear that his view does show, as he suggests it does, how it is that we can coherently conceive of impossibilities. For just as a view can imply a contradiction, without thereby showing how contradictions are possible, so can a view imply some state of affairs without thereby illuminating how that state of affairs is possible. So, how might it be possible that we coherently conceive of necessary falsehoods? If this is truly the case, and we are able to coherently conceive of

impossibilities, then why should we trust our concepts at all? They would seem to be, on such a view, wildly irrelevant to how the world is and might be. Loar does not pause to consider this, perhaps because he takes it that only triggered concepts, such as our phenomenal concepts, are stripped of their metaphysical implications. But more importantly, it seems impossible that we could coherently conceive of impossibilities. For suppose we can coherently conceive of some necessary falsehood. Then, since we are coherently conceiving of it, this means that in some possible world, that necessary falsehood is actually true. But there are no possible worlds in which a necessary falsehood is true. So we cannot conceive of necessary falsehoods.

Suppose Loar were claiming that the antiphysicalist can coherently conceive that phenomenal qualities are not physical-functional properties, even though the former just are the latter, by employing distinct names for each and believing these names referred to distinct entities. Just as someone can believe that Mark Twain is not Sam Clemens, so one can believe that phenomenal properties are not physical-functional properties. This is of no help; for the latter situation is not one in which someone coherently conceives of a situation in which Mark Twain is not Sam Clemens. Arguably, one can believe contradictions unawares; but this is not the same as coherently conceiving that Mark Twain exists even though Sam Clemens does not.

Could Loar, then, try to distance himself from this particular consequence of his view of phenomenal concepts? Could he rather present his view intact and simply claim that the anti-physicalists are wrong to think they can coherently conceive of the falsity of the a posteriori identity claim? I think that to do so would be tantamount to an extensive overhaul of his argument, if not his final view. For he claims to grant the anti-physicalist intuition of contingency, though it 'has been denied' by others, and go on to deny the general principle, which he takes to amount to the claim that our concepts have metaphysical implications.

2.221 The source of the incoherence: two sorts of cognitive independence

Where has Loar gone wrong? As suggested above, I think he has conflated (with White) two principles, one strong and false, the other weak and true. The strong principle is that if there is an a posteriori identity statement, then there are distinct properties that serve as modes of presentation of the object in question. This, I shall argue below, is false; and we can take Loar to have falsified this principle as well, with his above counterexample. But he takes himself to have falsified as well a weaker principle, which is that if we can coherently conceive of the contingency of an a posteriori identity statement, then there are distinct properties that serve of modes of presentation of the object(s). (There will be one object, plausibly, if the identity statement involves descriptions; but there will be two, if the conceivably false identity statement involves rigid designators.) This principle, I contend, is much more difficult to falsify in a substantive way. Even if we deny that there must be modes of presentation, rather than, say, some sort of direct reference relation, still there will be distinct properties, and these will moreover be of distinct objects, if the (conceivably false) identity statement involves rigid designators. For to genuinely conceive of the falsity of some alleged identity statement involving rigid designators, it is necessary that the identity statement actually be false. We cannot conceive of the contingency, and thereby falsity, of some genuine identity statement, though it certainly can seem that we can. To deny a principle like this weaker one would be to claim that what is conceivable can nevertheless be impossible; and this is what Loar cannot have done.

I think that on reflection it is apparent that Loar is making such a conflation. For recall his initial statement, that the PDA presupposes that 'the cognitive independence of *conceptions* ... implies the distinctness of the .... *properties* that those conceptions connote or stand for' (Loar, p. 2). In a certain light, I contend that this statement should strike us as difficult to deny. For if, according to our concepts, there is some distinction between properties, then is it not the case that those concepts imply that there is a distinction? Note that to imply that there is a distinction is not to make it

the case that there is one. On the other hand, if the statement is taken to mean that if our concepts imply that there is a distinction, then there is a distinction, then this should of course be rejected. For of course, this would rule out the possibility that our concepts are flawed in some way.

It would seem, then, that Loar is conflating, in statements like the above, two sorts of cognitive independence. On the one hand, there is the sort of cognitive independence that arises when we can find no a priori connection between two terms, or two concepts. Our concept of pain does not imply a priori that pain is a physical-functional property. That is to say, if pain is a physical-functional property, this is something we need to discover, through experiments on subjects in the actual world. Nothing follows from this sort of cognitive independence, because there are too many factors that could be barring the implication: perhaps our concepts are flawed, and if we settled on the right concepts there would be an a posteriori implication; perhaps pain is not a physical-functional property, and so there should be no such implication. On the other hand, there is the sort of cognitive independence that arises when we seem to be able to coherently conceive of the falsity of some a posteriori identification. If we can genuinely coherently conceive of the falsity of such an identification that involves the discovery that two descriptions are coreferential, then perhaps what we are conceiving is indeed some distinction in contingently related properties. On the other hand, if we are genuinely conceiving of the falsity of some alleged a posteriori identity statement that involves rigid designators, then the identification itself is false. (However, much work needs to be done before it becomes clear just which situations are genuinely coherently conceivable.) But now notice, that if Loar construes the general principle in terms of such intuitions of contingency, then it is wrong to say that no metaphysical implication follows from that fact alone. For they do. Furthermore, it is arguable that they follow even if the metaphysical implications are necessarily false, and the intuitions are of merely apparent contingency. Suppose that the antiphysicalist claims to imagine that the mental and physical are distinct. While the physicalist can see that

to imagine that the mental and physical are distinct. While the physicalist can see that what is claimed to be imagined is necessarily false, the antiphysicalist cannot see that; what is impossible can seem imaginable to some. Still what one imagines may on the most straightforward interpretation have the (necessarily false) implication that the mental is distinct from the physical. This does not mean that there is such a distinction (for the concepts employed may be partially incorrect), but it is certainly implied by the cognitive independence of the concepts, if cognitive independence is construed in terms of intuitions of contingency. On the other hand, suppose the intuitions are of the genuine contingency, and hence falsity, of the a posteriori identity statement. Then once more such intuitions have implications, and here they are necessarily true.

The conclusion, then, is that Loar cannot have falsified the weak version of the general principle, though it is plausible that he has falsified the strong version. But the real problem with the PDA involves the conflation itself, for it is this conflation which may seem to allow us to derive some metaphysical implication from the mere fact that there is no true a posteriori identity statement. In the next section, I will pinpoint the premise, in the argument for the general principle, at which White himself conflates the two sorts of cognitive independence. I will also go on to argue that the stronger version of the general principle is false, along lines similar to Loar's: the general principle assumes that the expressions in question cannot refer directly, but rather that they operate as descriptions. The thrust of my argument will be that while our concepts of phenomenal properties have numerous metaphysical implications (most likely not all of them true), I do not think that many have them are closely tied to the meaning of the relevant mentalistic expression. We may have mainly deeply ingrained beliefs, and these may be difficult to revise, but they nevertheless are not plausibly part of the meaning of the relevant terms. If this is the case, many implications of our concepts (expressed in the form of certain intuitions of contingency)

may be false, in that they may be plausibly seen as intuitions of merely apparent contingency.

3. What is wrong with the General Principle

Loar was right to challenge the general principle, and he was right to do so by bringing up direct reference. However, given his conflation of the two sorts of cognitive independence, he ended up claiming that what is conceivable may nevertheless be impossible. In this section I will argue that we can thankfully avoid that conclusion by bearing in mind the distinction between the two sorts of cognitive independence, and by focusing on the significance of the conceptual revision that might need to take place given certain a posteriori identifications. I will argue that White's general principle rules out the very plausible instance in which we are fairly ignorant about the ultimate nature of mental and physical properties, and as such should be wary of properties that we take to be given to us as modes of presentation. First I will present White's argument for the general principle. Then I will go on to discuss in more detail how I think it is false, given its conflation between two sorts of cognitive independence, and moreover its reliance on an implausible view of reference that is at odds with a direct reference view.

Why does White accept the general principle? In accepting the principle, White may be mistaking it for a weaker, more plausible principle that involves intuitions of genuine contingency. However, he does not spend much time explicitly discussing such intuitions of contingency, but only the a posteriori nature of the identification; the general principle is couched in terms of a posteriori identifications. To see exactly where the conflation between a posteriori identifications and intuitions of contingency is made, it is necessary to look more closely at the argument White offered in its support. White presents a reductio, on p. 92:

> Suppose that this [what is stated by the general principle] is not the case. Suppose, that is, that two descriptions are coreferential and that this fact cannot be established a priori and has not been established a posteriori.

> a possible world in which speakers who are epistemically equivalent to us use these terms to refer to different objects. There is, for example, a possible world in which the inhabitants are epistemically equivalent to those of our ancestors who used 'the morning star' and 'the evening star' before the discovery that the terms were coreferential and in which the inhabitants use the terms to refer to different planets. As used by the inhabitants of this possible world, these terms must pick out their referents in virtue of distinct properties because, unlike our terms, theirs pick out different objects. Hence the expressions as used by our ancestors must, contrary to our assumption, pick out their common referent in virtue of two logically distinct properties of that referent.
>
> The argument seems to be the following:

A) If there are two terms which are coreferential only a posteriori, then there is a possible world in which speakers epistemically equivalent to us use these terms to refer to distinct objects (or properties).

B) If speakers use two terms to refer to two objects, then these two terms refer via distinct properties of those objects (or properties).

C) If speakers epistemically equivalent to us use terms which refer via distinct properties, then when we use those terms, the terms refer via (those same) distinct properties — i.e. they have those same (distinct) senses in our world.

Therefore, D) If two terms are coreferential a posteriori, they have distinct senses and refer in virtue of distinct properties of the referent.

Loar seemed to question C, the premise that if speakers epistemically equivalent to us use terms which refer in virtue of distinct properties, then when we use those terms, the terms refer in virtue of those same distinct properties. As I see it, this premise is tantamount to the claim that our concepts, and what they suggest about other possible world, need have no implications for the actual world. For this premise states that if the meaning of our term 'pain', say, has some implication in some coherent situation, then it has that implication in our world as well. The point is that if we can conceive of some distinction, then that distinction holds in the actual world as well. (Suppose the distinction we conceive of is contingent, rather than necessary; e.g. the distinction between being Ben Franklin and being the inventor of bifocals. Then according to this

distinction we conceive of is contingent, rather than necessary; e.g. the distinction between being Ben Franklin and being the inventor of bifocals. Then according to this premise, the implication for our world is not that Ben Franklin did not invent bifocals, but that he might not have.) According to this premise, if something is genuinely conceivable, then this will have implications not merely for our concepts but for this world as well. On my view there is no difficulty with this premise.

I think the main worry concerns A, the premise that if there are two terms which are coreferential only a posteriori, then there is a possible world in which speakers epistemically equivalent to us use these terms to refer to distinct objects. For this premise amounts to the claim that whenever we have an a posteriori identification, we also have coherent intuitions that the identification could be false (if we make the plausible assumption that by 'these terms' he means that we must hold their meanings constant across worlds[4]); and this involves a conflation between the sort of cognitive independence that arises given an a posteriori identity statement, and that which arises given intuitions of contingency. Given this conflation, the premise may make sense for descriptions, but not for rigid designators, and so it is in this way that the principle the premise supports rules out direct reference. For the premise rules out the following possible situation: in our world, we discover that an identity statement is true. As such there are two terms that are coreferential a posteriori, say 'pain' and 'c-fiber stimulation'. However, there is no possible world in which speakers epistemically equivalent to us use these terms to refer to distinct referents, for pain just is identical to c-fiber stimulation. So, this premise does not hold true of terms that are directly referential rigid designators, as we should plausibly take 'pain' and 'c-fiber stimulation' to be. (Surely the descriptional appearance of expressions such as 'c-fiber stimulation' is not enough to render them non-rigid. Compare expressions such as 'H_2O'.) If we assume it holds true of all terms, including rigid designators, then we get the implausible

What of descriptions attached to the terms, 'pain' and 'c-fiber stimulation'? Surely we associate beliefs with both terms, especially the former. White may at this point try to object that it is unfair to assume that our terms do not rely on such beliefs, couched in descriptions, as senses. Then the argument would be that surely if we have distinct descriptions connected only a posteriori, this must mean that there is some possible world in which epistemically equivalent speakers use those terms to refer to distinct objects; and furthermore, there generally are such senses connected with our terms. At this point, then, the issue becomes whether there really are such descriptional senses attached to our mentalistic and physicalistic expressions, that we consider unrevisable. I think rather that while we do have beliefs intimately associated with these terms, it would not be wise to construe them as part of the sense of those terms, given that we are presently involved in constructing a theory of the brain and the mind. It may be, given our relative ignorance of the subject, that many pre-conceived notions will have to be refashioned.

To be clearer about what is at issue, we need to be clearer about what White takes sense, or modes of presentation, to be. He elucidates on p. 92: 'These modes of presentation of the object fall on the object's side of the language/world dichotomy. In other words they are aspects of the object in virtue of which our conceptual apparatus picks the object out; they are not aspects of that conceptual apparatus itself'. From this we can see that modes of presentation are construed as something more than properties we merely attribute to the referent, or think of the referent as having. They are rather construed of as properties the referent must have if we are able to refer to it at all. On a straightforward interpretation, these modes of presentation are given by, or recoverable from, the descriptive content of our concepts of the referents they present. So a mode of presentation on this view is a property of the referent that allows us to refer to that entity, by being the property that corresponds to the descriptional content of our concept of the referent. In order to refer to an entity by

referent that allows us to refer to that entity, by being the property that corresponds to the descriptional content of our concept of the referent. In order to refer to an entity by using a term, there must be (perhaps not in my thoughts at the time, but at least in the common language) some description associated with the term, and which *correctly* applies to the entity in question. These are the sort of reference-fixing properties that the general principle relies on; and such properties purport to tell us much more about the referent's nature than the properties that referent has solely as a member of an external relation.

I think there is room for suspicion about such a view (covered of course by Kripke in *Naming and Necessity*), especially when we consider the process of constructing theories about empirical phenomena. When someone engaged in the construction of some theory affirms an a posteriori identity statement, must we assume that there were distinct instantiated properties of the referent, given by earlier concepts of the referent? Couldn't those earlier concepts have rather not fully expressed a real property of the referent, while nevertheless managing to be concepts of the referent? They could manage to be concepts of the referent nevertheless by being at the end of some causal chain leading back to the referent; but this could be the case even if most if not all of the properties we were attributing to the referent given the descriptional content of our concepts were not in fact properties of the referent. The question, again, is not whether there often turn out to be properties of the referent which were presenting themselves to us, but whether the instantiation of these properties was a necessary condition for our ability to refer to the referents. We will not be able to generalize easily to other cases unless the existence of the properties is guaranteed in this way.

Consider the familiar Hesperus/Phosphorus example, upon which White relies. Were there really properties given by descriptions associated with the semantics of the names that enabled us to refer to Venus? There are at least two problems with

saying that there are. First, there is the problem of which descriptions in fact were taken to be part of the meanings of the names, granted that any such descriptions were. After all, the going concepts at the time were that Hesperus and Phosphorus were stars, not planets (let alone a single planet). (This is what prompts the redescription of the case to be that of one in which the reference is secured through descriptions like 'the celestial body that I see in the evening'.) Suppose, then, that the property of being a star was taken to be part of the meanings of 'Hesperus' and 'Phosphorus'. Was there, then, the property of being a star, which allowed us to refer to Venus? If a concept is a coherent one, then we may wish to say that there must be a corresponding property which is nevertheless uninstantiated;[5] but the real issue here is whether such a property must be instantiated, given the concept. If the property turns out to be uninstantiated, this is not enough for the proponent of the PDA, for the mental modes of presentation must be actual, not merely possible, for the conclusion to be interesting.

The proponent of the general principle could respond at this point that the property of being a star was not part of the meaning of the terms, although perhaps the speakers erroneously took this to be the case. Nevertheless, in the case of Venus, there was a relevant mode of presentation, or property, that was instantiated, namely that of looking like a star, or appearing to us to be a star. Why shouldn't we expect this sort of situation to generalize to other cases? Mustn't there be some mode of presentation in all cases that allows us to refer to the entity in question? We should not generalize, I want to suggest, because the same problem given above arises: not even the property of looking like a star was the referential mechanism that allowed our ancestors to refer to Venus; it was not genuinely part of the meaning. What allowed them (and us) to refer to that celestial body was rather the fact that they stood in some relation to it. This is made plausible by the fact that the description employed could in fact vary quite widely without its being the case that they were not referring at all. (Suppose they thought initially it was a space ship. Even so, weren't they managing however crudely to refer

to that object that is Venus, by misrepresenting it as a space ship?) The instantiation of any properties we attribute to the referent through the descriptional content of our concepts is not guaranteed as a matter of semantics. That is, no such properties need be instantiated in order for us to refer. Even where the beliefs we use to initially fix the referent are correct, and as such become firmly ingrained, this does not guarantee the instantiation of the properties in virtue of the meaning of the terms.

Perhaps a defender of the principle could object at this point that what he holds is not that the instantiation of the property of looking like a star, or anything that was explicitly in our concept at that time, was guaranteed by our ancestors' attributing that property to the referent. What shows that the property of being a star was not the true mode of presentation, or sense, of 'Hesperus' is just the fact that we were willing to revise such talk of stars out of the definition, in favor of the property of being a celestial body. These were our linguistic intuitions, and as such they reveal our semantics to us as we learn more about the world. If, however, we would not be able to stomach the claim that Hesperus and Phosphorus were not celestial bodies but rather, say, a space ship from Mars, this just goes to show the limits of our semantic commitments. (That is to say, were we to discover that what we call 'Venus' is in fact a space ship and its pilots have been cleverly deceiving us for quite some time, we would not put this by saying 'Venus has turned out to be a space ship'.) Still, the objector goes on, Venus had two distinct properties, one of being the celestial body seen at a particular place and time, and the other of being the celestial body seen at a different place and time; and so the general principle remains unfalsified.

However, we must continue to ask whether these two properties were ever guaranteed by our semantics. There is no philosophical dispute over whether Venus as a matter of fact presented such properties to us. The more difficult question is whether we could have referred, had such properties been absent; whether, that is, these properties were part of the meanings of the terms. And I think that if we give this

some thought, we will see that they were not part of the meanings. Even if we were to discover that the object we have been calling 'Venus' has been, all along, a space ship from Mars, we would still have been referring to some object, though we would have been wildly deceived about its nature; we would not put the discovery by saying that Venus has turned out to be a space ship, but there would be, I think, after this much time and causal contact with the object, a referential relation that would be held constant. Likewise, it is not part of the semantics of the terms 'Hesperus' and 'Phosphorus' that they name something that is seen at distinct places and times. We are not assured of the instantiation of these properties because they constitute the senses of the terms; we've become assured of their instantiation, rather, given the facts. We cannot perhaps easily imagine now revising some of these beliefs about Venus; but this does not mean that the initial star-gazer would have failed to refer had Venus not been presented to him in these ways. The beliefs may have initially helped to fix the reference, but they do not remain part of an unrevisable sense of the term. For suppose the star-gazer who first noted the position of the planet was confused about the time, or marked it down incorrectly in his logs. Still he could have referred to Venus, given merely his external relation to the planet. If this is plausible, then perhaps the most generalized description that can be applied to early uses of the names as part of their semantics is just that each names a celestial body. But the property of being a celestial body is a single property of the referent, and thus would not serve as an example of one of two distinct properties that served as the senses of descriptions that were discover to be coreferential a posteriori. It would seem, then, that this problem is a deep one for those who wish to uphold the general principle. Given Kripke's familiar points, the semantic theory upon which the general principle relies seems to have highly counter-intuitive consequences, and so should be denied.

One might object at this point that the above remarks are plausible enough, given that we have been talking about non-theoretical objects. In the case of

question is, how are we to relate the above points to the particular case concerning the nature of mental entities? First, I should grant that during construction of scientific theories, it will not always be a straightforward matter whether we are revising our concept of some entity on the one hand, or whether we are discovering the original posit simply does not exist. Nevertheless, there are some cases where it is fairly clear that what is going on is revision of some concept, that allows us to refer to the same entity we previously referred to, and forces us to conceive of it in a radically new light. The point to keep in mind is that in general, the instantiation of highly specific properties, given by descriptions we associate with the term, is not guaranteed as a matter of semantics. Their instantiation becomes in a sense guaranteed by the facts, if we become assured of their instantiation. But in general, making a posteriori identifications is not simply a matter of discovering that certain descriptions are coreferential; for those descriptions themselves will most likely be up for revision.

That is, as we are in the process of constructing theories, substantial reconceptualization is often required, and as such our concepts about the nature of the properties we attribute to the referents will get revised; so the implications we currently hold will themselves shift. While we begin by hypothesizing that certain descriptions are correct, thereby attributing properties to referents we take ourselves to be related to, the instantiation of these properties is not simply given by our semantic definition. Nevertheless, once the theory is more or less established, we do draw metaphysical conclusions from those conceptual definitions, on the assumption that we are in contact with the referent and that the referent does fit that description. (Here we can see how, contrary to Loar's view, the weaker version of the principle remains untouched: if our concepts dictate that an object has distinct properties, and as such we have intuitions that those properties could be instantiated separately, then we do, or should, believe that the object has distinct properties.)

that those properties could be instantiated separately, then we do, or should, believe that the object has distinct properties.)

Suppose, then, that we are wondering whether mental entities are identical to physical entities (e.g. neurological properties of the brain). Suppose further that according to the current concept of mental properties, these are essentially phenomenal, while according to the current concept of neurological properties, these are at least only accidentally phenomenal. As such, mental properties seem only contingently related, if at all, to physical properties of the brain; and these intuitions make it seem as though there must be at least distinct features being presented to us (if not distinct objects). We might even think that mental properties are necessarily only contingently related to physical properties of the brain. The property dualism argument would have us claim, given the general principle, that in identifying the referents of these two concepts, we realize that the referent must have been presented to us through distinct properties, one phenomenal, and one non-phenomenal, in order to account for the a posteriori nature of the identity. Given the above points, however, concerning the referential relation, there is at least one other path to take, that is, to realize that in our previous theorizing we thought a particular property was instantiated when in fact it was not: we wrongly attributed a property to the referent. Perhaps, for example, we were wrong to attribute the property of being only accidentally phenomenal to certain neurological properties.

It does not seem, in other words, very plausible to claim that it is part of the meaning of our mental terms that mental entities are not identical to physical properties of the brain. It may be a belief that is firmly ingrained in some, and it may be that such a belief played a large role in our first fixing the meanings of our mental terms, but nevertheless, such beliefs can be false. We can, that is, be wrong about what essential properties some referent has or lacks, and about the very nature of those properties. To suppose otherwise seems to be to confuse a metaphysical issue with a

conceptual issue. That some property is essential to an object does not mean that we must grasp that property with any certainty, or even that we fail to refer to the object if we fail to fully comprehend the nature of that property. This is not to say, naturally, that this path is the correct path to take in the case of mental entities. I have not provided an argument that shows that neurological properties are essentially phenomenal properties and vice versa. The point I have been making is that the alternate path, that we have been presented all along with the distinct mental and physical properties, is not forced upon us by the meanings of our terms. We could rather say that pain is identical to some neurological property of the brain, and is thus necessarily neurological; and that likewise, some neurological properties are necessarily phenomenal.[6]

Note that this way of denying the general principle is very different from Loar's way. On my view, while our concept of pain manages to be about pain in virtue of some external relation to the referent, and as such refers directly without higher-order reference-fixing, nevertheless, there is descriptional content associated with our concept of pain. This descriptional content we take to apply correctly to the referent (though of course we could be wrong), but we do not take this application to hold as a matter of the meaning of the term. We take it to hold because we think we are correct; however, we must remain open to reconceptualization. As such we must admit that there may be some aspects of deeply ingrained concepts that will have to change. Nevertheless, whatever concept we end up with, we will expect it to have metaphysical implications; and we will not take it that we can conceive of impossibilities. On this view, furthermore, we can also explain why it is that it seems possible to some that the relation between the mind and the brain is contingent, in a straightforward way: it seems this way to them, because (arguably) they have the wrong concepts, perhaps according to which physical properties of the brain cannot be essentially phenomenal.

These concepts are not simply triggered, however, and as such those who espouse the intuitions may eventually be open to reconceptualization.

Here is an objection to the above proposed view. According to our concept of pain, it is only contingently related to a neurological property of the brain. This is not just some belief we have about pain, akin to the belief that it is my Aunt's favorite mental property; it is part of what it is to be pain. Were we to discover otherwise, we would thereby claim that pain does not exist. Beliefs such as these are simply not up for revision. They are either held, or set free; and if the latter is the case, we conclude that we were not referring in the first place. But then the above story implies that we could discover that there is no pain, if our theory dictates that the corresponding definitive property we are trying to attribute to some mental entity is in fact uninstantiated. The point concerning reconceptualization then seems to amount to the claim that we could be radically mistaken about the fact that our experiences are essentially phenomenal, and as such it amounts to a radical eliminativism.

My reply to this is that, first of all, while there may be some beliefs that are simply not up for revision in this way, I see no reason to believe that the belief that pain is only contingently related to a neurological property is one of them. But furthermore, the objection takes my point about reconceptualization too narrowly. While I do take it to be part of the meaning of our phenomenal terms that they name phenomenal properties (whatever these turn to be), I did not take myself to be arguing that we must revise this meaning, and hence deny the phenomenality of phenomenal states. On my view it would take the plausibility of massive conceptual revision for eliminativism to become a reasonable option, regarding phenomenal properties. But in emphasizing the possibilities of reconceptualization, I did not mean to be claiming that our concepts of phenomenality need to bear the full burden. Rather, I maintained that both our concepts of phenomenality and non-phenomenality may need to change: we may need to reconsider whether what we typically considered to be non-phenomenal entities are

really non-phenomenal. We can, I repeat, be wrong about the essential nature of something, and still manage to refer to that entity. This does not mean that whenever we are radically confused about the essential nature of something, we still manage to refer. But in any particular case of such confusion, it is a nevertheless a possibility that we still manage to refer, and this possibility is one that is not ruled out a priori. Conceptual revision is complicated, and it need not always involve denying one conjunct of an apparent contradiction.

In sum, there are two problems with the general principle. One is that it conflates there being no true a posteriori identity statement with our having intuitions of genuine contingency; the other is that, while the former conflation may be unproblematic when it concerns non-rigid descriptions, it is implausible that we have unrevisable descriptions associated with our terms. The general principle rules out our being radically confused about the nature of our mental states, in that it rules out our mistakenly attributing phenomenal properties, as we presently conceive of these, to the referent, given that it interprets such properties to be unrevisable meanings. We must bear in mind that the possibility of our having been mistaken is a general possibility that has applied at various stages throughout our growing scientific knowledge.

4 Conclusion

In this paper, I have reconstructed a version of White's property dualism argument that does not beg the question against the physicalist, and that argued for the claim that unless there are expressions that are coreferential a priori with mentalistic expressions, there are irreducibly mental properties. The thrust of this reconstruction was that unless there are such a priori coreferential expressions, there will be an infinite regress of properties, which is impossible. I then considered an objection that there was nothing problematic in embracing such a regress; I replied that the regress is problematic in that it arises merely as a result of our semantic relation to the world. I then briefly examined Brian Loar's response to the PDA. He argued that its crucial

premise, viz. the general principle that if two expressions refer to the same entity and this fact cannot be established a priori, they refer in virtue of different modes of presentation of the referent, is false, in that it rules out the possible situation in which mental and physical expressions are coreferential only a posteriori, but the mental expression refers directly in virtue of a triggering mechanism, rather than through any mode of presentation of the referent. I argued in response that Loar inherits a conflation from White, between two sorts of cognitive independence, one involving a posteriori identifications, and the other involving intuitions of contingency. I claimed that Loar has presented a genuine counterexample to the general principle if it is construed as a principle that concerns a posteriori identifications; however, I argued that he cannot plausibly argue against the principle construed as one that concerns intuitions of contingency. Given his conflation, inherited from White, between our having no a priori connection between mentalistic and physicalistic terms and our being able to conceive that the relevant a posteriori identity statement is false, he ends up committed to the false claim that we can coherently imagine what is impossible.

At this stage, I moved on to present White's argument for the general principle, in order to argue more effectively against it. I argued against his claim that if there are two terms which are coreferential only a posteriori, then there is a possible world in which speakers epistemically equivalent to us use these terms to refer to distinct objects. It was here, I claimed, that White rules out the (plausibly actual) possibility of referring not through modes of presentation but through an external relation that we bear to the referent. While such a principle makes some sense if it concerns descriptions, it does not hold true of rigid designators; and furthermore, the descriptions we commonly associate with a given mentalistic expression should not be taken as part of the sense of that expression. What lies behind this criticism is in part a critique of a sort of Fregean view, according to which the referential relation is secured through modes of presentation, which are properties of the referent captured by our

current descriptional concepts of the referent. In effect my point is that such a view may be too simplistic, especially where theoretical construction is concerned, i.e. where we are trying to construct a scientific language which will enable us to explain certain phenomena. (Nevertheless, I grant that the above criticisms may speak against only an over-simplified Fregean view. My claim is that however one revises the view, a constraint should be that the revised view does not provide support for the general principle. For example, if a Fregean were to argue against me that she can incorporate all the benefits of the Kripkean points within her framework, I would be happy enough to accept such a view [barring any other problems with it] so long as it at least did not end up committing me to the general principle.)

In effect what the PDA and its general principle do, as espoused by White, is give undue metaphysical power to the fact that we have no a posteriori connection between mentalistic and physicalistic expressions, by adjoining this fact with the intuitions of contingency. While it is true that if we can coherently conceive of some distinction, then that distinction is possible, it is a difficult matter to establish just what is coherently conceivable. In particular, a posteriori identifications need not always give rise to such coherent conceivabilities. Given, then, the mistakes that ensue from the general principle, which provides the heart of the PDA, we can safely say that the argument is unsound, and so does not provide a compelling argument for the truth of property dualism. In the next chapter, we will see what happens when the intuitions of contingency, construed as merely epistemic, are deprived of their metaphysical significance.

Chapter 2

Never Mind the Gap

0. Introduction

Some philosophers of mind, among them Joseph Levine and Colin McGinn, argue that the mind-body problem is solvable in a metaphysical sense but is subject to a further, epistemological problem known as 'the explanatory gap'. That is, while the physicalist's answer to the metaphysical question 'What is the relation between the mind and the body', viz., that the relation is that of identity, is most likely correct, it nevertheless fails to explain why it is that a certain mental experience, e.g. a sensation of pain, feels the way it does rather than some other way. In other words, the physicalist's theory fails to make it transparent why it is that pain, as some physical property of the brain, must feel the way it does. That it fails to do so is shown by what are known as "intuitions of contingency": some feeling of contingency seems to linger stubbornly around any identification of the mind with the body. That is, it seems as though it is possible for me to be in the same physical state that I am in now, while my experiential state is entirely different, and vice versa. Furthermore, while other philosophers (among them Saul Kripke, Thomas Nagel, and John Searle) claim that these intuitions pose a problem for physicalism as a metaphysical thesis, proponents of the gap claim that these intuitions of contingency do not speak to the metaphysical mind-body problem itself, but rather serve only to contribute to the above epistemological problem. The explanatory gap is taken to be merely epistemological, impotent against the truth of the physicalist's hypothesis as a metaphysical thesis. Nevertheless it is posed as a serious problem for physicalism.

What would bridge the explanatory gap, according to its proponents, would be some transparently necessary statement relating the mind to the body; however, the physicalist's hypothesis of identity does not strike them in this way, and furthermore they think it never will so strike them. No matter how much evidence is

garnered in support of the truth of physicalism, the claim is that it will never support the explanatory power of physicalism. As such, the physicalist's view that mental entities simply are physical entities allegedly fails to make them fully intelligible to us (the presumption being that the intuitions are widespread). According to the proponents of the explanatory gap, we feel we are told in a sense what the entities are; but we are mystified about why they are thereby experienced as they are.

In this paper, I will question the notion of the explanatory gap as a merely epistemological problem for physicalism that leaves the truth of the metaphysical thesis untouched. Ultimately, I do not think the gap as it is presented poses any epistemological problem for physicalism that is separable from a metaphysical difficulty for the view: any reasons given for thinking there is an epistemological gap will serve also to cast doubt — however weak or dispersible — on the metaphysical thesis that mental entities are physical entities. I do not claim to prove this beyond any reasonable doubt, but only to arouse some serious suspicions. The questions that I intend to deal with are: What is the nature of the alleged gap? How is its existence argued for? Do these arguments fail to apply to the metaphysical thesis of physicalism? and finally, Is the gap a legitimate concern for physicalists? My central claim will be that one prominent way of arguing for the explanatory gap collapses into a challenge to the metaphysical thesis, and should be treated as such.

1. The nature of the gap

The intended nature of the explanatory gap can only become clear after examining arguments for it, but we can do something to discuss what the shared conclusion of those arguments is supposed to be. What is the epistemological problem for physicalism that is meant to be distinct from any objection to the truth of the metaphysical thesis? Roughly, the problem is this: given the identification of mental (phenomenal) states[7] and any state that seems definable without mention of essential phenomenality (for example functional states, physical states, purely intentional states,

or some combination of these), we are not able to determine from that identification why it is that the mental state has the macro-properties it has, i.e. why it feels the way it does rather than some other way. So, for example, consider the physicalist's identification of pain with c-fiber stimulation. As an identity statement, it must be necessary (assuming, as I will for the course of this paper, that the terms of the relevant sentence expressing the statement are rigid designators). Nevertheless, we (may) have intuitions that this identification is contingent — that pain might not have been c-fiber stimulation, but some other sort of property of the brain, or perhaps some non-physical property. Similarly, we seem to be able to imagine that c-fiber stimulation could exist, even though the subject would experience no pain, or perhaps even no sensation whatsoever. (Analogous intuitions can arise for a supervenience theory.)

It might seem to some that the implication of such intuitions is that physicalism, as an identity thesis or as a supervenience thesis, is false. However, such intuitions, given that they are intuitions about what is conceivable, point only to an epistemic possibility, so the argument goes, and as such do not count against the metaphysical thesis of physicalism. This epistemic possibility, or conceivability, nevertheless poses a serious problem for physicalism. For, while the intuitions do not imply that physicalism is false, they do show that physicalism does not make the nature of mental properties fully intelligible to us. According to physicalism, mental properties are physical properties of the brain, but this merely describes the mental properties; it does not explain how it is that they are as they are, at the macro-level. For why should c-fiber stimulation feel the way pain feels? We know only that c-fiber stimulation happens to feel painful, even if we are convinced of the truth of physicalism; we do not, so the proponent of the gap contends, have a genuine explanation of why this is the case. To put the issue another way: what would satisfy those who mind the gap? What would bridge the gap would be our being able to see how the (phenomenal) macro-properties of mental entities follow from, i.e. are

determined by, the fact that those entities are physical properties of the brain. This is what is allegedly left unexplained by physicalism.

I take it that if this is the case as Levine says, and that if the identification of mental entities with physical properties of the brain does not and could not bridge the explanatory gap, then this is indeed a concern for physicalism. For suppose one tried to object that it is unfair to demand that physicalism explain why mental properties feel as they do; it's enough, rather, that physicalism pinpoint the underlying, physical nature of the properties, i.e. it's enough to describe the mental properties as physical. I do not think that this would be a satisfactory line to take. For surely, barring the extreme view of eliminativism, it is a fact to be explained that phenomenal properties do feel a certain way, that pains feel different from tickles, and so on. The phenomenal, experiential nature of certain mental properties — that they feel a certain way — is an interesting fact, and it is reasonable to demand its explanation. A theory that did not provide such an explanation, or even the hope of one, would be a less than satisfactory theory.

However, the proponents of the gap have a large task, which is to show that physicalism is unable to provide even the hope of some future explanation of the phenomenality of certain mental states. It may be that, given merely the identification of pain with c-fiber stimulation, it is not yet obvious how this explains why pain feels as it does. Nevertheless, we must ask, why is it taken by the proponents of the gap to be a foregone conclusion that physicalism will never be able to explain this phenomenality? Perhaps, that is, further theorizing is needed; perhaps we need to see what differentiates c-fiber stimulation from a-, b- and d-fiber stimulation, before we can say more about why it is that pain feels as it does rather than as a tickle. The proponents of the gap, in particular, need to beware of merely claiming at each step in our theorizing 'but why does that explain why pain feels as it does?' For we need some reason to think that this skepticism (skepticism, that is, of the truth the proposition that pain feels as it does

because it is some particular physical property of the brain) is plausible. Furthermore, the proponent of the gap needs some reason which does not count as an objection against physicalism as a metaphysical thesis. If the quizzical attitude is meant as an expression of epistemological puzzlement, then this puzzlement cannot simply collapse into the skepticism of a frustrating and curious child who will accept no explanation of the color of the sky. In a certain mode, we can always get ourselves to feel puzzlement over why a proposition is true; but there must be more to the explanatory gap than this, for it to be a substantive challenge. Whether there is some substantive challenge is what we shall now go on to consider.

2. The arguments for the gap

In this section, I present McGinn's and Levine's arguments, interestingly different from each other, for the explanatory gap. First I discuss McGinn's view, and argue that the diagnosis of the flaw in his argument has to do with a false dilemma that he proposes, that there are only two ways we could come to understand the mind-body problem. He claims that we could only come by a theory of consciousness by studying the brain, or by introspecting consciousness, and that neither of these (perhaps not surprisingly) will work. Once we recognize that this dilemma is a false one, we can realize that McGinn's argument is not compelling. I then go on to present Levine's argument, which replies explicitly on a Kripkean argument involving intuitions of contingency. Here I argue that Levine gives us no compelling reason to think that the intuitions of contingency are as stubborn as he thinks they are. (They might turn out to be stubborn; but this possibility is not enough for us to resign ourselves to the fact that they will turn out to be stubborn.) I then move on in section 3 to discuss whether it is really the case that the gap, on either view, is distinct from a metaphysical challenge.

2.1 McGinn's argument for the gap

In McGinn's words, 'we are cut off by our very cognitive constitution from achieving a conception of that natural property of the brain (or of consciousness) that

accounts for the psychophysical link' (McGinn, pp. 2-3). His conclusion is, then, that there is some naturalistic answer to the mind-body problem, but that we are cognitively closed to this answer, given our methods of concept-formation. What exactly does it mean to be cognitively closed to a theory? As he puts it, 'A type of mind M is cognitively closed with respect to a property P (or theory T) if and only if the concept-forming procedures at M's disposal cannot extend to a grasp of P (or an understanding of T)'[8] (McGinn 1990, p. 3). What is it, then, to grasp a property? McGinn is unclear here, but presumably to grasp a property involves some deep, non-superficial understanding of its nature. This is still vague, but we can say a little more. Presumably, to be merely in perceptual contact with a property is not sufficient to grasp it; we need to know something substantial about its role in producing the relevant phenomenon, to be able to grasp it. We need, that is, to arrive at and understand the correct theory of its nature; and this is what we cannot do, in the case of consciousness.

McGinn's argument for the claim that we cannot understand the answer to the mind-body problem goes as follows (see McGinn 1990, pp. 7-14):

M1. Consciousness is a natural phenomenon.

M2. So there is a natural property of the brain, P, that explains consciousness — i.e., the instantiation of which determines the presence of consciousness.

M3. But we have only two concept-forming procedures which could lead us to a grasp of P: the procedure that involves i) the direct study of consciousness, which would involve either introspective phenomenology or conceptual analysis; or ii) the study of the brain.

M4. We cannot grasp P by engaging in introspective analysis.

M5. If we can grasp P by studying the brain, then either P is graspable by perception alone, or the existence of P is legitimately inferable as an explanation of some properties that are graspable by perception alone.

M6. We cannot grasp P by perception alone.

M7. That P exists is not legitimately inferable as an explanation of properties graspable by perception alone.

M8. So our concept-forming procedures cannot lead us to grasp P.

Therefore, we cannot grasp P, i.e. we are cognitively closed to P, although we can be reasonably sure that P does exist.

Note that McGinn's conclusion could be interpreted in various ways of various strengths, depending on how we interpret 'concept-forming procedures at M's disposal'. First, what exactly does he mean by 'concept-forming procedures'? McGinn is not explicit here; he tells us that 'minds are biological products like bodies, and like bodies they come in different shapes and sizes, more or less capacious, more or less suited to certain cognitive tasks' (McGinn 1990, p. 3). He goes on here to contrast cognitive closure with perceptual closure, noting that there would seem to be clear cases of the latter: the perceptual capacities of humans are quite different from those of a mollusk, say. So, the analogy might be, just as he we have the five senses, we have a number of cognitive procedures. But what are these? All we have to go on are McGinn's two proposals for how we might come to learn about consciousness: by introspecting consciousness, or by studying the brain. I don't think he means to be making some general claim about types of cognitive abilities. We have perhaps all sorts of cognitive abilities, which allow us to learn about numerous sorts of things; for example, there are cognitive abilities that allow one to learn about music, there are abilities that help one to learn about mathematics, about language, about flying airplanes, and so on. Why then should we think that there are only two procedures, when it comes to learning about the nature of consciousness? I will get into this in more detail below, but I think perhaps what he has in mind is relatively simple: we seem to have two sorts of data. One sort involves our subjective, conscious experience, and the other involves the study of the brain and its neurological structure. Suppose, as is likely, that there is some link between these. Shouldn't we be able to learn about this connection by studying either one, or the other? If there is some connection, then it seems we should be able to happen upon it by studying either relata, for either relata

will be adjoined to the connecting link. Consider the following crude analogy: suppose we are trying to learn about the nature of an actual bridge between two islands. Shouldn't we be able to happen upon that bridge by learning all we can about either one island, or the other? For in either field study, we shall happen upon the bridge itself. But now, suppose, the bridge is made of some material that our senses cannot take in, or that is so complex and foreign to us that we would not conceptualize it as a bridge. I think that this is what he has in mind; and as we shall see, this relatively naive idea is the weak point of his argument.

What concept-forming procedures are relevant, then, will depend on the particular field of study. If we are trying to learn about music, there will be various procedures to employ, for example learning to play various instruments, learning musical theory, learning to memorize songs, and so on; but if we are trying to learn about consciousness, the claim is, there are only two that are currently at the disposal of our type of mind. Which brings us to the next question for McGinn: how are we to interpret 'at M's disposal'? Are we to take this to mean merely 'currently at M's disposal'? Certainly not, for this interpretation would allow that we could develop, in the near future, some concept-forming procedures that would allow us to understand the answer to the mind-body problem; this is an uncontroversial conclusion, and McGinn is aiming for something more exciting. Moreover, his account of cognitive closure is a claim about types of minds, and as such its content will also depend on whether we presume that all humans have the same type of mind. For example, could certain neuroscientists grasp the answer to the mind-body problem sometime in the near future, while lay people would remain mostly in the dark?

Call the following the strongest interpretation of cognitive closure: A type of mind M is cognitively closed with respect to P if and only if the concept-forming procedures presently at M's disposal cannot extend to a grasp of P, and it is *logically impossible* for M to form the appropriate procedures that would extend to a grasp of P.

At first glance, we might think that this is the conclusion McGinn is aiming for, for he does say things like 'I do not believe we can ever specify what it is about the brain that is responsible for consciousness' (McGinn 1990, p. 2); he seems to want a strong, "in principle" conclusion. Furthermore, it seems that he has given us an a priori argument for our never being able to explain consciousness; why should we think it possible at all to do so? For if it is logically impossible for us to understand the answer to the mind-body problem, how could it be possible for anyone else? Nevertheless, this strong conclusion would be in tension with his emphasis that his argument concerns our cognitive condition; we are 'constitutionally ignorant at precisely the spot where the answer exists' (McGinn 1990, p. 21). And McGinn rightly distinguishes his conclusion from that which would go hand in hand with the strongest interpretation of cognitive closure, by claiming that perhaps the answer to the mind-body problem is only, as he says, 'relatively closed' (McGinn 1990, pp. 15-16) to us; perhaps some mind could come to grasp it, in particular if that mind had other concept-forming procedures (whatever they might be).[9] Perhaps, in fact, even we as a species could come to grasp it — but this would involve our evolution into a species radically different from our present form.[10] Call this the strong interpretation of cognitive closure: a type of mind M is cognitively closed with respect to P if and only if the concept-forming procedures at M's disposal cannot (logically) extend to a grasp of P, and there is no way for M to form the appropriate concept-forming procedures that would extend to a grasp of P without at the same time effecting a significant change in the species of animal that has the type of mind M. In other words, this sort of development, leading to a grasp of P, would be tantamount to a profound development of the species, which would involve a radical difference in the type of mind that is typical to the species. An example of this sort of cognitive limitation would be the way in which a horse, say, cannot speak a language. A talking horse wouldn't be a horse at all, and its mind would not be the mind of a

horse. This strong interpretation, I take it, is the appropriate interpretation of cognitive closure for McGinn's purposes.

Weaker interpretations of the principle would be illustrated by examples such as a 5-year-old's present inability to grasp the theory of relativity. That 5-year-old does not now have, and may never go on to develop, the appropriate concept-forming procedures, involving mathematical concepts and procedures, and if he does he may simply never use them; but we would still want to say that they are in some way at his disposal. They are there for him to develop, if he is so inclined. Weaker interpretations such as these, however, are clearly not strong enough for McGinn. For his conclusion is that we humans, and not merely we laymen, or even we humans at this stage of our knowledge about the mind, cannot understand the solution to the mind-body problem. That we need some conceptual innovation before solving the problem is, while not agreed on by all sides, much less controversial than McGinn's conclusion.

I think that we should be wary of the strength of McGinn's conclusion. It seems that in his eagerness to avoid the dogmatic claim that we can understand anything whatsoever, given our scientific abilities, he has put forward an equally dogmatic claim that we can at this moment know that there is something that we cannot ever understand: a particular explanation of a particular natural phenomenon. He rightly notes that 'the human mind ... must conform to *some* principles — and it is a substantive claim that these principles permit the solution of every problem we can formulate or sense' (McGinn 1990, p. 5). That this is a substantive claim is admittedly true. Fortunately, though, to attempt to formulate a theory of the mind, we need not rely on a principle like this; we need only rely on the principle that we've got some reason to hope that we can indeed come to understand most natural phenomena, and so we've got some hope to think that we will eventually be able to understand the mind. At any rate, it is also a substantive claim that the human mind must conform to the same principles at all points in its evolution, as McGinn appears to believe; and it is,

moreover, false. McGinn's example of the Humean mind and its puzzlement at the physical world (which has subsequently evaporated, at least for those of us schooled in physics) is actually a rather nice example of the possibility of the evolution of the principles of the human mind, without any corresponding evolution of the species. Are we willing to call ourselves of a different species than that of David Hume? We can agree that perhaps we are now to the mental world as Hume was then to the physical world, but we should hesitate to agree that we will always be in such a condition. We may grant that it is possible that we will never understand the solution to the mind-body problem, for it is possible that there are things we could not, or will not, ever understand, and that this is one. But this weak claim applies to any puzzle whatsoever. What is so special about mental phenomena, according to McGinn, that should make us think that some more significant claim applies? The answer lies in McGinn's premise M3.

2.11 Assessing McGinn's argument: a false dilemma and its effects

There are, I think, many problems with McGinn's argument, and so there are many points at which one might challenge it. Owen Flanagan, in Chapter 6 of *Consciousness Reconsidered*, contends that the main fault with the argument lies in its premises M5 - M7; he claims that McGinn is construing our methods of theorizing about the brain too narrowly. I agree that McGinn is guilty of this, as we shall see in more detail below, but I think the real problem lies in the crucial premise M3, according to which we have only two ways to discover the correct explanation of consciousness, i.e. to discover why it is as it is and why it exists at all: we can either focus on the brain, or introspect the phenomenology of conscious experience. McGinn provides no explicit support for this premise. However, we can see why he might find it somewhat plausible. Recall that the nature of the explanatory gap is supposed to be such that we can never see, or comprehend, why a phenomenal state feels the way it does, even if we have identified it with a brain state. It just does not seem determined by the

identification of pain with c-fiber stimulation that pain should feel like pain, rather than a tickle, or an itch, or nothing at all. For why does c-fiber feel the way it does? As McGinn puts it, 'How could the aggregation of millions of individually insentient neurons generate subjective awareness? ... Somehow, we feel, the water of the physical brain is turned into the wine of consciousness, but we draw a total blank on the nature of this conversion' (McGinn 1990, p. 1). There seems, intuitively, to be no necessary connection between c-fiber stimulation and any sort of phenomenal experience. For McGinn, this demand (coupled perhaps with the naive idea that we should be able to learn about the connection between any two things by simply learning all there is to know about one or the other) translates into the demand that we be able to either perceive or infer, from studying the brain, the existence and very nature of phenomenal states; or that we be able to introspect or infer the nature of the brain states from the existence of the phenomenal states. We cannot, it seems, either perceive directly or infer indirectly, by studying the brain, anything about the nature of the phenomenal state that the brain is in, or whether it is in one at all. So, it seems, it will never strike us as determined that the brain should be in some phenomenal state rather than another, or none at all. Even if it is somehow metaphysically determined, it will never seem to be determined from our standpoint.

But note that the demand has at this point become rather dubious. It is one thing to demand that, given a theory, the existence and nature of the relevant macro-properties are determined, and seem so to us. If, given the identification of water with H_2O, and given our background knowledge of chemistry and physics, it still seemed ad hoc that water should behave at the macro-level just as that stuff in the oceans does, we would rightly feel puzzled. (Moreover, we would rightly, it seems, question the acceptability of the theory that identified water with H_2O molecules. This is a separate problem for the proponents of the gap, to be discussed further below: how do the arguments manage to leave the metaphysical thesis untouched?) But why

should we ever think that we could have come to understand the relation between water and H_2O molecules, by studying either only the macro-properties, or only the micro-properties? If we never looked at both sorts of phenomena together, simultaneously, we never would have identified water with H_2O in the first place. Likewise, it is overly restrictive to claim that we must, prior to having a theory fairly well worked out, stare long and hard at the brain, and be able to deduce what phenomenal properties are being instantiated, if we never allow ourselves to study both sorts of phenomena at once. There is, to be sure, always a puzzle in studying a phenomenon the nature of which one does not fully understand. How are we able to latch onto the nature of the phenomenon, or even assure ourselves we are referring to the right thing, given that we do not understand that nature? But this is a puzzle that pervades scientific theorizing, and in fact all learning, and as such is not sufficient to compel us to throw our hands up regarding the mind-body problem.

Once we relinquish the false dilemma that is premise M3, however, it becomes unclear why we should not suppose that the appropriate concept-forming procedures are somewhere and sometime within our reach. What we need to do is somehow study both consciousness and phenomenal experiences along with what we know about the brain and its properties, to see if we can identify the link between the mind and the body. Further problems with McGinn's argument, then, arise as a result of this false dilemma. The most interesting of these, I think is what happens when he tries to support premises M5 - M7, in which he claims that we cannot legitimately posit or infer the existence of consciousness by studying the brain alone. McGinn argues that we cannot recognize the link between mind and brain by simply perceiving (looking at) the brain; nor can we approach it through any sort of legitimate inference (McGinn 1990, p. 11). I take the first claim to be trivially true, if we do not yet have a theory in hand, and probably false if we do have a theory, but I won't go further into it here. The more problematic claim is that we cannot infer the explanation of consciousness. Why

should we suppose that we cannot infer the nature of the property that links the mind and brain, by studying the brain alone? McGinn claims that

> a certain principle of homogeneity operates in our introduction of theoretical concepts on the basis of observation ... [C]onsciousness itself could not be introduced simply on the basis of what we observe about the brain and its physical effects. If our data, arrived at by perception of the brain, do not include anything that brings in conscious states, then the theoretical properties we need to explain these data will not include conscious states either. (1990, pp. 12-13)

He goes on here to argue that we have no working model for introducing the theoretical concepts that would apply to the link between the mind and the brain:

> ... [W]e arrive at the concept of a molecule by taking our perceptual representations of macroscopic objects and conceiving of smaller scale objects of the same general kind. This method seems to work well enough for unobservable material objects, but it will not help in arriving at P since analogical extensions of the entities we observe in the brain are precisely as hopeless as the original entities were as solutions to the mind-body problem. We would need a method that left the base of observational properties behind in a much more radical way. ... [N]o concept needed to explain the workings of the physical world will suffice to explain how the physical world produces consciousness. (1990, p. 13)

So, the claim is that given the methodological principle of homogeneity, we are not allowed to infer the existence of any property that somehow goes beyond what is needed to explain the observed data. The pressing question for McGinn is how are we to decide what 'goes beyond' what is needed to explain what we observe?

In other words, what exactly is the content of the principle of homogeneity, and why should we adhere to it? Flanagan suggests that McGinn is here relying on Nagel's claim, which McGinn quotes, that 'it will never be legitimate to infer, as a theoretical explanation of physical phenomena alone, a property that includes or implies the consciousness of its subject' (Nagel 1979, p. 183). McGinn is, I think, relying on this idea; but Nagel's claim cannot itself provide support for the principle, since it is too close to the principle itself. Flanagan also argues here that McGinn has forgotten that both brain facts and facts about consciousness are to explained, and so clearly any

homogeneity constraint that disallows us to introduce consciousness will be too strict. But at this point, I think there is a larger criticism to be made, namely that if we agree with McGinn's third premise that there are only two ways to discover the answer to the mind-body problem, we have already granted him too much. The relevant criticism then becomes that the strictness of the homogeneity principle is flowing from the faulty reasoning behind the third premise. Criticizing premise M7 for its disallowing the introduction of conscious properties is to sidestep the real issue.

Flanagan comes closer, I think, to a deep criticism of the homogeneity principle when he discusses how we might infer the existence of electrons:

> Given a commitment to standard contemporary physics, it is the inference to the best explanation that certain observable processes in a cloud chamber are the traces of unobservable electrons. We never see the electrons directly while observing the processes in the cloud chamber, nor for that matter do we see them anywhere else. Electrons are a theoretical construct whose postulation best explains certain observable data and whose postulation is in turn supported by certain (predicted) observations. (p. 113)

Here, Flanagan's point seems to be that it is not a problem for theory construction in general that we have to infer the existence of objects or properties that we do not directly observe. But in this he seems to be misinterpreting the homogeneity principle: he seems to be taking the principle to be one which does not allow us to make an inference from the observed data to unobserved theoretical constructs. As he says on p. 114, '[the principle] is untenable if it is meant to render it impermissible to draw explanatory links between some set of events of processes unless all these events or processes are simultaneously observable in the domain under study'. Such a consequence of the principle would indeed serve to cast doubt on the principle. But this interpretation of the homogeneity principle is itself too restrictive. McGinn himself thinks that inferences from observations of physical data to theoretical (physical) constructs are legitimate, since what is being postulated as an explanation of the data is not 'different in kind' from the properties observable in the data: some analogies, e.g.

those that allow us to infer the existence of molecules, are legitimate, even given the homogeneity principle. So the phrase 'different in kind' seems to be crucial in arriving at an appropriate understanding of the principle.

The principle, then, seems to be the constraint that when we infer from our data that unobservable properties or objects exist, those properties or objects must be of the same general kind as the properties and objects that we observe immediately in our data — i.e., as the 'observational properties'. What McGinn seems to be saying is that the inferences we typically make about the existence and nature of unobservable entities are constrained, in that we must conceptualize the nature of the latter as being straightforwardly analogous to the nature of what we perceive directly at the macroscopic level in the everyday world. In this way, we make no wild departures in our inferences about atoms that do not seem to fit with what we observe in the macroscopic world; we do not conclude that their structure and nature is unlike anything that we can see around us. However, when it comes to mental phenomena, so the argument goes, it seems there is such a large distinction between what we experience as consciousness and what we find at the level of the goings-on in the brain, whatever inference we would make would indeed involve a radical departure from the typical sorts of things we can directly observe. There seems to be no available model for inferring the existence of some property that accounts for the link between two sorts of things so disparate as the mind and the brain. This, I take it, is the force of the principle of homogeneity: the phenomena in question are so seemingly heterogeneous and different in nature that there just is no plausible analogy that would help us see how postulating some unobservable property would explain why that property explained the correlation between the observable property of the brain and the relevant subjective, phenomenal property.

While this interpretation of the principle allows it to be more complicated and more accurately aimed, it is nevertheless too strong to be plausible. It is not clear

that the inferred properties of atoms and sub-atomic particles are as straightforwardly analogous to the properties of the macroscopic world as McGinn would have us believe. It remains obscure how we are to decide which properties are similar enough in kind to the macroscopic ones as to be left untouched by the principle. This point is related to that of Flanagan's: what we observe, after all, are processes in cloud chambers. Exactly how are the electrons that we go on to postulate as a result similar in kind to these processes? Or to any other processes and objects that we observe in the macroscopic world? One analogy that was once popular was that atoms were rather like a solar system, with a sun and its surrounding planets. However, I think it should be clear that an analogy such as this will not take a theory very far. The explanation of the origin of a planet, for example, will be rather different from the explanation of the origin of an electron; and the same goes for the explanations of their respective behaviors. This sort of model provides merely a (sometimes) useful heuristic. Furthermore, there are some posited entities — or forces — which seem to have no analogy whatsoever. Consider, for example, the force of gravity. What in the macroscopic, every day world does this resemble, so that we are allowed to infer its existence? It seems that theoretical inferences are an extremely complicated matter. So, while we may at this point have no appropriate model for explaining the link between the mind and the brain, this is not due to the fact that no such model is possible given the homogeneity principle. In fact it is unclear that the homogeneity principle in any plausibly true form would be able to rule out such an analogy; and interpreted strongly, it cannot plausibly be applied to other legitimate inferences. So why should we hastily apply it to the case of mental phenomena?

The answer is that such a principle, interpreted strongly enough so as to rule out a legitimate inference in the case of mental phenomena, will seem plausible to those who think that we should be able to perceive the phenomenal state of a subject, if we are given merely his brain state, before we have a theory in hand. That is, it will

seem plausible to those who think that premise M3 is true, and that we are not allowed to consider information about the subject's phenomenal state along with which neurons are firing. For if we are allowed both of these sorts of data at once, and if we go on to theorize that phenomenal states are brain states, then there is no problem about having to infer which phenomenal state is going on from a description of the subject in solely neurological terms. So the plausibility of the homogeneity principle again seems to arise from the false dilemma we encountered in premise M3. [11, 12]

While more could be said about McGinn's view, I think we can learn more about the alleged explanatory gap by moving on to Levine's view and seeing how it is similar to and different from McGinn's.

2.2 Levine's argument for the explanatory gap

Levine approaches the gap by way of arguments that are purported to be challenges to the metaphysical claim made by physicalism, that the mind is the body: Kripke's argument concerning intuitions of contingency, and Jackson's argument concerning what Mary didn't know. It is important to consider these arguments, since Levine claims (and presumably McGinn would agree) that they do not constitute a real challenge to the metaphysical view of physicalism, but rather serve to support an epistemological point, since they reveal the inadequacy of physicalism as a genuine explanation in its inability to make the mental intelligible to us. Crucial to his argument that intuitions support this epistemological claim is that the intuitions will never subside. My main criticism of Levine is two-fold: first, if we agree that physicalism is likely to be true, there is at present no compelling reason to think that the intuitions will never subside; and furthermore, and most importantly, the intuitions themselves do constitute a metaphysical challenge to the truth of physicalism, and were they never to subside we ought to have metaphysical worries. Indeed it is puzzling, if the intuitions have no metaphysical implications, as Levine contends, how it is that he thinks he can draw other conclusions about physicalism from them.

First, we need to look more closely at the arguments that are put forward as arguments against the truth of physicalism. I will focus, for simplicity, on what Levine says about Kripke's argument. Roughly, Kripke's argument is as follows (see Kripke 1972, pp. 144-155): given an identity statement, professed to be true by the physicalist, such as the statement that pain is c-fiber stimulation, we nevertheless have intuitions that the statement could be false. We seem to be able to describe coherent situations in which c-fiber stimulation is going on, but there is no experience of pain. We also seem to be able to describe situations in which there is pain, but no c-fiber stimulation. However, given that the identity statement is between two rigid designators[13], it must be necessarily true if it is true at all. So, there should be some way of explaining away the appearance of contingency, if physicalism is true, as we can with other seemingly contingent statements like the claim that heat is molecular kinetic energy. Kripke contends that this latter statement as well may seem contingent, but such intuitions can be explained away as being more correctly describable as intuitions of the following sort of situation: something other than molecular kinetic energy affects us in the way that it happens to affect us in the actual world, i.e. what we imagine (and what is indeed possible) is that some phenomenon other than heat feels warm to us. But this is just a world in which molecular kinetic energy either does not exist, or affects us in some way other than it does in this world; it is not a world in which molecular kinetic energy is not heat. Kripke maintains that there is no obviously similar way to explain away the feeling of contingency about the identification of pain with c-fiber stimulation, since to be in pain just is to feel pain, and vice versa. I cannot coherently maintain that I could be in pain, and yet not feel pain. So, the conclusion is, that until physicalism can find some way to explain these intuitions away, its truth stands challenged.

I think that there are other (perhaps less philosophically interesting) ways to answer that challenge, which Kripke does not consider. I think the physicalist

should respond that those intuitions of contingency are wrong, or, that is, are intuitions of merely apparent contingency, and that once we have more of a theory in hand, we will be able to see that the identification is indeed necessary (whatever the identification turns out to be). It is simply too early at this stage of theorizing to suppose that those intuitions are very trustworthy (though we certainly can examine them further and see what sort of view might develop were we to take them as genuine intuitions). However, I will not try to argue for this further here. What I do wish to note here is that this sort of reply grants that the intuitions do have metaphysical implications (more on this below), but that we will in the future be able to see why those implications are false, and thus that we should reject the intuitions as pointing to a merely apparent possibility.

Levine, on the other hand, claims that these intuitions do not challenge the metaphysical thesis, because as he says, given Kripke's own distinction between metaphysical and epistemological possibility, 'metaphysical consequences cannot be drawn from considerations of what is merely conceivable' (Levine, p. 124). Apparently, they are intuitions merely of how things seem, not of how things genuinely might be. As such, the intuitions support only an epistemological claim, that even though physicalism is probably true, we will never be able to see it as a satisfactory explanation; physicalism does not make mental phenomena intelligible to us, because the intuitions will never abate. In Levine's words, the argument is as follows:

> [T]here is also an epistemological sense of 'leave something out', and it is in this sense that conceivability arguments, being epistemological in nature, can reveal a deep inadequacy in physicalist theories of mind.
>
> For a physicalist theory to be successful, it is not only necessary that it provide a physical description for mental states and properties, but also that it provide an explanation of these states and properties. In particular, we want an explanation of why when we occupy certain physico-functional states we experience qualitative character of the sort we do. ... [W]hat is at issue is the ability to explain qualitative character itself; why it is like what it is like to see red or feel pain. ...

> The basic idea is that a reduction should explain what is reduced, and the way we tell whether this has been accomplished is to see whether the phenomenon to be reduced is epistemologically necessitated by the reducing phenomenon, i.e. whether we can see why, given the facts cited in the reduction, things must be the way they seem on the surface. I claim that we have this with the chemical theory of water but not with a physical or functional theory of qualia. The robustness of the absent and inverted qualia intuitions is testimony to this lack of explanatory import. (1993, pp. 127-129)

From these remarks, the argument seems to be the following:

L1. When we have an explanation of a phenomenon, e.g. when we identify water with H_2O and analyze our concepts of water in terms of chemical concepts, the phenomenon that is reduced is "epistemologically necessitated" by the reducing phenomenon (i.e., its existence and thereby the instantiation of its properties, will be and will seem to us to be determined given the existence of the reducing phenomenon and its properties).

L2. Given the robustness of the intuitions of contingency of a statement such as the claim that pain is c-fiber stimulation, mental phenomena are not epistemologically necessitated by the proposed reducing phenomena.

L3. No matter how detailed the physical-functional theory of mental phenomena gets, the robustness of the intuitions of contingency will remain. [That is, not only will the intuitions themselves remain, but they will remain robust: we will never find the means to explain them away.]

L4. Physicalism (in the form of either an identity thesis, or a supervenience thesis) is most likely true.

L5. The intuitions mentioned in L2 and L3 do not speak against the truth of physicalism; they constitute no metaphysical challenge to physicalism.

Therefore, physicalism does not, and cannot, provide an explanation of mental phenomena, though it is probably true.

In the next section, I will argue that there is tension between L3 and L4, which arises from the falsity of L5. What I shall argue in response to Levine in this section is that L3 is probably false, for the same reason that it would be false were it used in a metaphysical challenge to physicalism: the intuitions are too pretheoretical to be heavily relied on. L3 rests on the claim that there is a disanalogy between the statement expressed by 'water is H_2O' and that expressed by 'pain is c-fiber

stimulation'. Is this disanalogy really as striking and unbridgeable as Levine suggests? Why should we think that L3 is true?

Levine claims that no matter how much theory we built up concerning the brain, it would still seem legitimate to wonder why pain feels as it does, given that it is c-fiber stimulation. But how plausible is this premise? If we are in a skeptical mode, could we not do the same for water, and certain of its macro-properties? Why is it, after all, that water freezes at one particular temperature rather than another, given that it is H_2O? I confess that the links between the macro- and micro- properties of water seem contingent to me; we might say that I have my own explanatory gap, between H_2O and its macro-properties. Presumably this is because I do not know enough about chemistry. This is how Levine answers the worry, in fact: 'If someone asks why the motion of molecules plays the physical role it does, one can properly reply that an understanding of chemistry and physics is all that is needed to answer that question ' (Levine 1983, p. 358). So the response to my explanatory gap between H_2O and its macro-properties is that I am ignorant. Fair enough. But the relevant point is that presently we are ignorant about the full nature of the brain, and as such intuitions of contingency should not be relied on too heavily. If the response works in the situation where someone is ignorant of an established theory, it should also work in the situation in which someone is ignorant given that there is no established theory.

But someone might worry that what is being imagined in the case of phenomenal properties is really slightly different. After all, Levine claims to be relying on the intuition that pain seems only contingently related to c-fiber stimulation, even given the plausibility of the identification as put forward by physicalism; I, on the other hand, cannot claim to be coherently imagining that water is something other than H_2O molecules. I must say that I cannot coherently describe such a situation, since, as we all know, water is H_2O. But I can seem to imagine it: I say to myself 'this stuff I am drinking is really H_5O'. As it turns out, what I seem to be imagining is not a coherent

possibility. I can imagine it having turned out that something like water was H_5O; but I cannot imagine that water is H_5O. Likewise, it can seem coherent to me that something other than H_2O has its essential macro-properties. This is easier to describe: couldn't H_5O appear just as water appears to me? The point is that what can seem perfectly imaginable from a standpoint of ignorance (be this a standpoint of pre-theoretical ignorance, or a case of simple ignorance of an established theory), can nevertheless be impossible.

Suppose someone objects that there is still a distinction between my seeming to imagine that water is not H_2O, and my seeming to imagine that pain is not c-fiber stimulation; for in the former I can coherently explain what it is that I am imagining. However, in the case of pain, I cannot claim that I am able to imagine something other than pain affecting me the way pain does. Pain just is the unpleasant effect. I would reply that the explanation only makes sense given the established theory that water is H_2O. If there were no such theory available, then even though I could not explain away the seeming coherence of the imagined scenario, the imagined scenario would nevertheless be necessarily false. But let us grant that the intuitions of the contingency of the relation between mind and brain are not the same as the intuitions concerning water. What is truly analogous to Levine's explanatory gap is the gap between granting that water is H_2O, and seeing transparently that H_2O must have the macro-properties that it in fact has. Levine treats pain as a supervening macro-property of c-fiber stimulation. Consider a section like the following on p. 129 (1993):

> What is explained by the theory that water is H_2O? Well, as an instance ... let's take its boiling point at sea level. ... I claim that given a sufficiently rich elaboration of the story, it is inconceivable that H_2O should not boil at 212 degrees Fahrenheit at sea level. But now contrast this ... with a physical or functional reduction of some conscious sensory state. No matter how rich the information processing or the neurophysiological story gets, it still seems quite coherent to imagine that all that should be going on without there being anything it's like to undergo the states in question.

It seems coherent to imagine no pain even though there is c-fiber stimulation; but this is only if we are ignoring the claim that pain is c-fiber stimulation. I cannot say: 'pain just is c-fiber stimulation, although they are coherently distinct.' If, on the other hand, I take pain to be a macro-property of c-fiber stimulation, then just as I can be ignorant about the way in which the macro-properties of water are determined, I can be ignorant about how the macro-properties of c-fiber stimulation are determined. The relevant point here is that it seems imaginable, from a standpoint of ignorance, that H_2O should have different macro-properties, and yet it does not constitute any sort of interesting criticism (metaphysical or epistemological) against the theory that water is H_2O, and thereby has certain macro-properties necessarily, precisely because it is put forward from my ignorant point of view. What seems imaginable from a standpoint of ignorance may turn out to be not only false but necessarily so.

Furthermore, given that we are at an early stage of theorizing about the mind and the brain, how can we predict that these intuitions will never subside, and never be able to be explained away? If the intuitions really do remain robust, then this might be some problem for physicalism (though it would be, I shall argue below, a metaphysical problem); but the proponent of the gap must claim that it is fairly obvious right now that they never will subside, or be explainable in some way. (Perhaps they will remain as some foggy illusion that can be dispelled once we apprise ourselves of the theory; perhaps I could dispel my intuitions that H_2O could have different macro-properties, had I enough time and interest to learn the details of chemistry.) However, given the fact that presently there is no fully worked out theory of the mind and its relation to the brain, I do not see how this important demand can be met. So it seems the best response on behalf of the physicalist is to challenge Levine's premise L3: there will be no substantial explanatory gap, once we have enough of the theory worked out. It will then be determined, and seem determined, given merely the fact that a subject is in a certain brain state, that he or she is experiencing a phenomenal state.

In fact, it seems that an identity theorist could say that we need not wait for further theorizing; it should be seen as determined right now, if we think that pain is c-fiber stimulation. If I as a physicalist hold that pain is c-fiber stimulation, then if I am told that Fred's c-fibers are being stimulated, I will conclude that Fred is in pain. If Levine challenges me to explain why pain feels as it does rather than feeling like a tickle, I can respond that c-fiber stimulation is an essentially painful state, while d-fiber stimulation (the brain state correlated with tickles, say) is not. For if we identify pain with c-fiber stimulation, and we are asked why c-fiber stimulation is painful, we could simply say 'What do you mean? It just is pain.' If we are pressed to explain why it is that pain does not feel like a tickle, the answer is that c-fiber stimulation is not d-fiber stimulation, and they have rather important differences (assuming, on the basis that physicalism is plausibly true, that we will eventually discover such differences). So perhaps some of the intuitions of contingency arise, then, because we simply do not yet know which states those phenomenal states are; we do not know which brain state constitutes pain, and which a ticklish feeling. But once we have worked this out, it will be determined (metaphysically) and seem determined (epistemologically) that c-fiber stimulation feels the way it feels, rather than as a tickle, because of the differences between c- and d-fiber stimulation.

Suppose Levine presses on that this response does not answer his worry that the feeling of contingency will be stubborn. Won't the intuitions remain, in that we will still be able to imagine that the phenomenal aspects of c- and d- fibers are reversed? Note that by this he must mean more than simply the fact that the intuitions will remain, even though we will see how it is that physicalism falsifies them. For if we do eventually come up with some difficult explanation of why it is that pain feels as it does rather than some other way, and if nevertheless some people still had such intuitions, the appropriate reply would be simply that they fail to understand the explanation. This is what Levine must mean by the 'robustness' of the intuitions. For if the

robustness merely involves the fact that some will go on having them regardless of the explanation, these will not suffice to throw doubt on the power of the explanation; they will rather be akin to the child's endless curiouity about what explains the color of the sky.

So Levine must mean that right now, we know that the intuitions, though plausibly false, given the truth of physicalism, will never be able to be explained away, and they will never subside even in those who are fully schooled in the final theory. My reply is that this worry is premature. The intuitions are too pre-theoretical to be much of a pressing challenge. It may be that they will remain; but I do not see that one should be sure of this right now. What sort of support can be given for them, and for their alleged robustness? The most obvious answers to this seem to amount to a mere restatement of the intuitions themselves: "phenomenal experiences just seem so unlike brain states". Furthermore, bear in mind that these are intuitions that someone who accepts the plausibility of physicalism is meant to have. It is true that we do not have the identifications completely mapped out; but were we to have a theory worked out to fuller detail, according to which each phenomenal state is identical to some state of the brain, it would be puzzling why someone would stubbornly resist the identity as explanatory. After all, in making the identification we are not denying that pain is essentially painful. We are simply saying that we have discovered that pain is also essentially a state of c-fiber stimulation, just as we discovered that heat is essentially molecular kinetic energy. (It may be that in such a case, there were conceptual connections that helped the identification strike us as plausible. And so, conceptual revision may be required before the a posteriori identification of the mental with the physical seems plausible. But this is just to say that eventually, it is likely that the intuitions of contingency will either subside or be explainable.) Similar remarks, it seems, would follow for a supervenience sort of physicalism.

And after all: is a view like dualism better off, in respect to the explanatory gap? It seems that no matter what we identify pain with, even if we just say that it is made of (essentially painful) mental stuff, we could always demand to know why that stuff feels as it does. That it is part of the meaning of 'pain' does not explain, after all, why pain feels as it does, any more than the meaning of 'bachelor' explains why bachelors exemplify the true nature of bachelorhood. Why are those fellows, and not others, unmarried? The dualist can reply that it feels painful rather than ticklish, because it's a pain and not a tickle — but the physicalist can say the same thing. In fact, the physicalist can say more: that painful feeling is c-fiber stimulation, and as such has many differences from d-fiber stimulation. But if the intuitions can be wielded against any view at all, they become less of a substantial worry and closer to a radical skepticism.

Is there some other way to explain why someone might think that the intuitions will remain robust? Perhaps what is underlying the stubborn feeling of contingency is the sort of intuition we have when we claim that trees just are the molecules that make them up. This doesn't seem necessarily true, for after all, given a particular tree we can lop off a chunk of its matter and nevertheless the tree will survive. So the tree is not strictly identical to the stuff that composes it; and we go on to make a distinction between the 'is' of identity, and the 'is' of composition. But these intuitions, applied to pain and the stimulation of c-fibers, while they may have some consequences for our view of the relation between the mind and the body, are not the sort of intuitions that someone like Levine has in mind. He has in mind intuitions that pain could arise as a result of the stimulation of d-fibers, or some radically different sort of stuff, perhaps. And again, such intuitions may simply reflect my current state of ignorance, and perhaps should be revised once we learn more about the facts. At the moment we are all ignorant of the details of what sorts of states mental states are; but this just goes to show that intuitions about their being only contingently related to states

like c-fiber stimulation cannot carry much weight. This is not to say that they carry no weight; but they do not constitute a substantial challenge to the explanatory power of physicalism. Given the pre-theoretical nature of the intuitions, we have no compelling reason for thinking that L3 is true and that the intuitions will remain no matter how much evidence for physicalism we garner. (Again, the intuitions might remain. I do not mean to rule out this possibility.)

Notice that my reply to Levine's explanatory gap, an allegedly merely epistemological trouble, seems to be the same reply that I briefly mentioned above as a response to Kripke, who took the challenge to be one against the metaphysical thesis of physicalism. This may seem puzzling, if we think that the gap is separable from the metaphysical challenge. At this point, I will go on to discuss Levine's reasons for thinking the explanatory gap is 'merely' an epistemological problem, and whether this is indeed the case. My thesis will be that the worries that allegedly lead to the gap, namely the intuitions of contingency, should be treated as nothing over and above the original metaphysical challenge. Levine is wrong to claim that one needs an argument to get from epistemological possibility to metaphysical possibility. Cases of genuine (as opposed to apparent) conceivability do have metaphysical consequences. The real issue, then, lies in what is truly conceivable, not whether our intuitions have metaphysical consequences.

3. Conceivability and metaphysical challenges

We have seen above that the explanatory gap seems in this way to be indistinguishable from a metaphysical challenge to physicalism: the response to the intuitions construed epistemically is the same response to the intuitions construed metaphysically. Why is it that, according to both Levine and McGinn, the gap is not taken to threaten the truth of physicalism? The answer to this will depend on exactly how the gap has been argued for. On McGinn's view, recall that we are supposedly cognitively closed from the answer to the mind-body problem because we allegedly have only two ways of learning the answer, and neither way will work. On this sort of

have only two ways of learning the answer, and neither way will work. On this sort of view, it is plausible to see that the truth of physicalism is not obviously threatened. Physicalism may be true; we will just never know that it is true, given the way our minds work. However, note that McGinn's argument for the claim that there is some naturalistic answer to the mind body problem is exceedingly weak. He claims that there must be some answer, since we do not think that consciousness is a magical phenomenon. It must have some explanation, and McGinn takes this to mean that there must be some naturalistic explanation. But if we agree with him that we have only two ways of finding out the answer, we may wonder why we should think that there is answer — let alone any answer that approaches physicalism. (It is perfectly consistent that there is an answer that is out of our reach; but this is true for any scientific question, and the more difficult question is why should we think it applies with any profound significance to the case of the mind-body problem?) After all, we have no way of finding out what the full answer is, and so it would seem to be a matter of faith whether we believe that there is an answer or not. By which concept-forming procedure do we conclude that there are no magical, brutely unexplainable phenomena? So ultimately McGinn's view, depending on where we put the emphasis, may have some bite against the metaphysical thesis of physicalism. For, were the argument sound, it would effectively bar us from ever having positive reason for finding physicalism plausibly true (or from finding any other theory of the mind plausibly true). This does not constitute a challenge to the truth of any view: it does not amount to the claim, e.g., that physicalism must be false because it entails falsehoods. But it is a challenge to anyone who thinks that it is reasonable to think that physicalism is true. For if we are convinced of our cognitive limitations, what basis do we have for joining in McGinn's move from premise M1, that consciousness is a natural phenomenon, to M2, that there is a natural property of the brain that explains consciousness? Why should we even believe premise M1?

How Levine's view allegedly manages to avoid constituting a metaphysical threat to physicalism depends on the support given for its explicit premise L5, that the intuitions of contingency do not support a metaphysical point. On first sight, the metaphysical modesty of the view may seem clear: physicalism may be true, but we will never find its truth completely transparent, or satisfying, given our stubborn epistemic intuitions of how things seem. However, I think that the explanatory gap as argued for by Levine ultimately collapses into the earlier metaphysical challenge. Here I will argue first, if L3 is true then L4 is false; and if L4 is true, then L3 is false. There is a weaker version of L3, which allows that the intuitions of contingency may be merely apparent intuitions; if we reinterpret them in this way, if L3* is true, we have prima facie reason to doubt the truth of L4*; and if L4* is true, we have prima facie reason, barring a global skepticism, to doubt the truth of L3*. Furthermore, given this tension between L3 and L4 (on either construal), L5 is highly problematic, regardless of his argument for it. I will then go on to show his argument for L5 is not compelling.

First, why is there tension between L3 and L4? The reason for this involves Levine's support of L3. Why is it that such intuitions exist? According to Levine, this is because 'our concepts of qualitative character do not represent, at least in terms of their psychological contents, causal roles' (Levine, 1993, p. 134). For if our concepts of phenomenal mental properties purported to be of properties that were definable by their physical-functional causal role, then there simply would be no corresponding intuition that we could distinguish the mental property from an abstractly specified causal role. We would not find it at all surprising, say, that the Chinese people as a group could, in virtue of being in a certain functional organization, be a singular conscious entity (see Block). The reduction of one entity to another is explanatory, i.e., epistemologically necessitates the existence of the former given the latter, when we are able to specify some defining causal role of the former and go on to

analyze the concepts of the one in terms of the other. But then since reductions are explanatory only by way of specifying the causal role of the object or property to be reduced, any object or property not fully constituted by its causal role will not be explained by reduction.

However, this elucidation of the argument should strike us as puzzling, as it enhances the tension between L3 and L4. For what are the intuitions of contingency, exactly? They seem to be intuitions that the identity thesis is false; i.e. they are intuitions, arising from our concept of pain, that pain could exist without the relevant physical property, and vice versa. But if these are the intuitions, and we grant that they are coherent, it would seem that they do imply the falsity of the identity thesis. They are a counterexample to any claim that the mind is necessarily related to the brain. But then this is in direct conflict with L4. Likewise, if physicalism is true, then the relation between the mind and the brain is not contingent, and so the intuitions cannot be successful, or genuine, intuitions of contingency. As such if physicalism is true, then we have no *genuine* intuitions of contingency. Levine claims that 'the main burden of the physicalist argument is borne by considerations of causal interaction. If qualia aren't physical processes ... it becomes very difficult to understand how they can play a causal role in both the production of behaviour and the fixation of perceptual belief' (Levine 1993, p. 126). This may be so. But why should this consideration carry more weight than (hypothetically) coherent intuitions that pain is distinct from c-fiber stimulation, i.e. that physicalism is false? Further, this revealing quote emphasizes an important point, that if it is true that according to our concept of phenomenal properties, they do not occupy a causal role, or there is some feature of them that does not occupy a causal role, then this is in conflict with any claim that they do occupy a causal role, or that they wholly occupy a causal role. It is obscure how the intuitions get around their business of having metaphysical implications.

In other words, given this tension between L3 and L4, it seems that the intuitions of contingency that figure in premises L2 and L3 do have metaphysical implications; that is, L5 seems false. For could we not run a similar argument with very different, metaphysical connotations, in the following way: phenomenal properties do not occupy causal roles, or at least their full nature does not; reductions only explain those objects or properties that do occupy causal roles; so phenomenal properties cannot be explained by reduction; and moreover, they cannot be *identified* with a causal role, given the first premise. Why should the elucidation commit Levine to the first premise, that phenomenal properties do not occupy causal roles? Because, as he says, our concept of phenomenal properties does not represent a causal role. So, according to our concept of phenomenal properties, they are not causal roles, or at least there is some feature of them that is not simply a caual role. But, a supporter of Levine may add, perhaps we are simply wrong, and our concept of phenomenal properties is not applying to the sort of thing that we take it to be applying to. If this is the case, I reply, then why do not the intuitions have (incorrect) metaphysical import? To say that one has a false or flawed concept is not to say that the concept has no metaphysical implications. It is rather to say that the concept has false metaphysical implications; the concept is not matching up appropriately with the world.

It seems, then, putting aside for the moment Levine's argument for L5, that the premise is highly problematic, since if our concepts are such that they represent phenomenal properties as non-physical entities, then this representation seems to amount to a metaphysical implication. Another way to put the point is the following: suppose that phenomenal states do not occupy a physical-functional role (and our concepts are correct on this score). Does this not speak against the truth of the metaphysical claim that phenomenal properties are physical-functional properties? On the other hand, suppose that our concepts of phenomenal properties are such that they imply that the properties are not identical to physical-functional properties, since our

concept of them implies that they do not merely occupy a physical-functional role. Are we to be satisfied with the claim that physicalism, though out of synch with our concepts, nevertheless is true? It seems that it would make more sense for us, in such a situation, to either revise our concepts, or to revise physicalism, or at least our belief in it. In short, it seems that the argument, given its crucial reliance on the intuitions of contingency, is hard pressed to avoid constituting a challenge to the metaphysical claim made by physicalism, premise L5 notwithstanding. To get clearer on this issue, we need to look more deeply into Levine's reasoning behind his premise L5, in which he relies on a distinction between epistemological and metaphysical possibility.

First, however, it may seem to some that I am misinterpreting Levine's argument, and in particular his conception of the intuitions of contingency. Suppose he does not take it that the intuitions of contingency are genuine intuitions of a coherent possibility. Suppose he takes the intuitions to be of an apparent contingency, which may or may not turn out to be genuine. Then we should reinterpret L3 to be

L3* No matter how detailed the physical-functional theory of mental phenomena gets, the robustness of the intuitions of apparent contingency will remain

At this point, one might think that here is the true explanatory gap: no matter how much we learn about the brain and the mind, it will always seem that they are only contingently related, even if our theory dictates that those very intuitions of contingency are merely apparent. Still, however, I think there is a tension between L3* and L4. For suppose L3* is true; the apparent intuitions of contingency remain unexplainable to the very end. Why should we take it that they are intuitions of merely apparent contingency, rather than intuitions of the real thing? I take it that generally, if we have intuitions that a statement is contingent, and we can find no means after much theorizing to explain such intuitions away, this would count as a reason for supposing that the intuitions were genuine. In such a case, of course, it would remain possible that nevertheless our concepts are a faulty guide, and that while they seem by all accounts to

be of a genuinely possible situation, they are not. But this is something, if one is in a Cartesian skeptical mode, that applies to any concept of ours. As such it poses no special problem for our concepts of the mind. On the other hand, suppose that physicalism is most likely true; if we find this plausible, what reason do we have presently for supposing that the intuitions of apparent contingency will remain apparent, regardless of all the evidence we garner? If we truly believe that physicalism is true, then it would seem likely that eventually either the intuitions would subside or we would find some way to explain them away as being intuitions of merely apparent contingency. This is not to say that it is incoherent to suppose otherwise; but again it is always coherent to suppose that our concepts, even given all the evidence, are not applying to what we take them to apply to. This is just to say that all the evidence does not entail the truth of our theories. The evidential relation is not one of deducibility.

Levine has a reply to the above, which is that still I am misinterpreting his view of the intuitions. On his view, it is wrong, or misguided, to claim that the intuitions are of apparent contingency, leaving it open that they may be of genuine contingency. This is misguided, because intuitions, as he says, are merely an epistemological matter. As such, we need not take them to be any guide to the world as it really is; they have no metaphysical implications. If they have no metaphysical implications, then clearly they make no challenge against the truth of the physicalist's thesis, of which we are assured of on other grounds. How, then, does Levine claim that the intuitions themselves do not constitute a metaphysical challenge to physicalism? How does he argue for L5? His argument is a negative one, inspired by Kripke's claim that the intuitions do support a metaphysical point. Consider Kripke's argument, as discussed earlier. There are two things that might be said, in reply to it: one is that to have Cartesian intuitions with Kripke is to make a false metaphysical point; and another is that these intuitions never even succeed in making a metaphysical point. Levine seems to want the latter. As he says,

> [W]hat is imaginable is an *epistemological* matter, and therefore what imagining pain with c-fibers does is establish the *epistemological* possibility that pain is not identical with the firing of c-fibers. It takes another argument to get from the epistemological possibility that pain is not the firing of c-fibers to the metaphysical possibility, which is what you need to show that pain isn't in fact identical to the firing of c-fibers. (1993 p. 123)

He also says:

> ...a consequence of their [i.e. the early identity theorists'] theory is that it is not possible for some mental state not to be identical to its physical or functional correlate. But the basis of Kripke's objection lies in a strict distinction between metaphysical and epistemological possibility. Once we appreciate that distinction, the physicalist can return to her original ploy, i.e. to say that metaphysical consequences cannot be drawn from considerations of what is merely conceivable. (1993 p. 124)

And in an earlier paper he writes:

> Since epistemological possibility is not sufficient for metaphysical possibility, the fact that what is intuitively contingent turns out to be metaphysically necessary should not bother us terribly. It's to be expected. (1983 p. 356)

I agree with the claim that it is not very surprising or troublesome that 'what is intuitively contingent turns out to be metaphysically necessary'. However, I beg to differ with the reasoning behind it. (Epistemological) intuitions of contingency do have metaphysical implications. This is shown by the fact that once we learn more about how we think the world in fact is, we will either see that there was something wrong with those earlier intuitions, or we will see that the intuitions were in fact pointing to a real possibility. If the situation is the former, we won't go around espousing them any more (or eventually, even being struck by them), once we see that they were not matching up with the facts as we expected them to. Levine, on the other hand, seems to want to deny in passages like the above that the intuitions of contingency have any metaphysical implications whatsoever: they are "merely" epistemological. It is as if he can have such intuitions and say simultaneously 'this is how it seems to me, but I don't think this is how it really is;'[14] The correct view, I think, is that if we have such

intuitions, these intuitions tell us how the world, out there, seems to be. In other words, they have metaphysical implications. They show us a little bit about what our concepts are; and we further think that we have roughly the right concept, though we do not rule out revising it. But given that we take our concepts to represent the world, we expect to be able to draw (cautiously) metaphysical conclusions from them.

My response to Levine, then, is that surely our concepts have metaphysical implications. One might think that Levine could not be denying this, and that I am interpreting him unfairly. But consider, in contrast to my point, the following passage from Levine:

> [S]uppose we reject the Cartesian model of epistemic access to metaphysical reality altogether. One's ideas can be as clear and distinct as you like, and nevertheless not correspond to what is in fact possible. The world is structured in a certain way, and there is no guarantee that our ideas will correspond appropriately. If one follows this line of though, then the distinction Kripke points out between the pain/C-fibres case and the water/H_2O case turns out to be irrelevant to the question of what is or is not metaphysically possible. (1993, p. 123)

Levine seems to be suggesting that if we admit that our concepts are fallible, then there is simply no trusting them at all. Ironically, he takes himself to be construing Kripke as someone who follows the later meditations in which we refute the skeptic by having clear and distinct intuitions; but he himself seems to favor the position of the equally implausible Cartesian skeptic. For the intuitions we derive from our concepts, on this view, turn out to be *irrelevant* to the way the world is. The demand that we give an argument from what is conceivable to what is metaphysically possible translates into the demand that we refute the skeptic: what we can conceive, genuinely or apparently, has no bearing on the metaphysical possibilities. But notice that if we need such an argument connecting genuine conceivability with possibility, then so does Levine, for we are all in the same situation. If his claim is correct, it turns out that we have no reason to believe anything whatsoever, given our concepts. This may not seem obvious, but consider the force behind the statement that our concepts

have no implications whatsoever: if what he means is that we could always be employing the wrong concept, and that this shows that we can draw no metaphysical conclusions from our concepts, then we cannot draw metaphysical conclusions from even our most firmly established concepts. It is always consistent that our concepts are wrong; there is no guarantee that the world is as our concepts suggest. But this does not in general deprive them of their metaphysical implications. In fact, what it is for them to be wrong is to have false metaphysical implications.

He seems, then, to fail to see the strength and generality of his claim that nothing metaphysical follows from what is conceivable. What he does in effect is give in to skepticism, for our theories are built partially out of what is conceivable, and even more modestly, out of what we take to be conceivable. Even if we grant that the intuitions may be only of an apparent possibility, still they are not irrelevant to the way the world is, that is, they are not without implications. Of course they are irrelevant if what we mean by this is simply that they might be wrong. But why should we conclude from the fact that concepts are irrelevant in sense, that we can draw no metaphysical conclusions from the concept whatsoever? If there is some compelling reason to believe that the concept is the wrong concept, then this would be a reason to withhold the metaphysical implications, not because the concept has none, but because the implications are, on further reflection, false. Likewise, consider the Cartesian skeptic: if the point is merely that it is logically possible that all one's concepts are wrong (and I do not mean to be suggesting that this is all there is to Descartes' skeptical argument), why should this make us conclude that we can draw no conclusions from our concepts? If the explanatory gap is directly on a par with this sort of skepticism, unbeknownst to its proponents, then it is not the direct challenge it has been taken to be. For the explanatory gap began as a problem peculiar to the mystery of the mind. If it expands into skepticism, it should not be seen as a particular worry for philosophers of mind who find the identity thesis plausible.[15]

Contrary to Levine, then, we do not need an argument from the claim that something is (genuinely) conceivable to the claim that it is metaphysically possible. Nor do we need an argument from apparent conceivability to genuine conceivability; for this would be to refute the skeptic. What we do need, however, is some fairly compelling reason to think that what is apparently conceivable is genuinely conceivable, if we want to go ahead and draw the implications. This is what Kripke cannot supply, given the pre-theoretical nature of the intuitions; whether the situation is genuinely conceivable will depend on what the ultimate nature of the mind is. Kripke, however, does manage to appreciate that in general this is where the problem lies: in distinguishing what seems conceivable from what really is conceivable and therefore metaphysically possible. It is crucial to his argument, in fact, that we might think we are able to imagine something, e.g. a distinction between heat and molecular kinetic energy, that we cannot imagine, since it is metaphysically impossible not to think we imagine it, but to succeed in imagining it. Levine's problem, however, is not that Kripke holds that whatever we imagine willy nilly is metaphysically possible. His worry is rather that nothing metaphysical follows from claims of conceivability, and that we need an argument to establish otherwise.

No such an argument is required, however. It may still seem to some that this is too quick: for after all, if I think that something is imaginable, shouldn't I be careful in drawing a substantive conclusion from it? To this I answer, yes; but the reason is not that our concepts have no implications. The reason is quite the opposite: our concepts do have implications, and there are an awful lot of false ones, and not many true ones. Rather, what we must be careful about is whether what seems conceivable really is conceivable. To be assured of this, however, unfortunately a lot of metaphysics must already have been settled. But we do have to start somewhere, and hence we do need to rely on our concepts to draw metaphysical conclusions. After all, we need something to draw the conclusions from. Recall that Levine's worry is not that

Kripke draws the implication too quickly; it is apparently that he draws the implication at all.

So my point against Levine is really twofold: first, what is genuinely conceivable does have metaphysical implications; and second, what *seems* genuinely conceivable also has metaphysical implications, and these are implications that we can and do draw cautiously. We do not draw them cautiously because we are not sure whether the concept really has that implication; we draw them cautiously because we are not sure whether we are employing the right concept, with the correct implications. The reason for caution is that those implications themselves may turn out to be false (perhaps even necessarily so). Nevertheless, we must put some faith in our concepts, for whether we decide to settle on the concepts we presently have or to refashion them or relinquish them altogether, we will hopefully plan to draw (roughly correct) metaphysical implications from whatever concepts we end up with. (For this reason it is no scathing objection to Kripke's argument that he takes the concept he is working with to be roughly correct.) For what route could there be to metaphysical possibility, apart from epistemological possibility, sufficiently considered and examined? It is puzzling how we could somehow stumble upon metaphysical possibility without couching it in terms of what is conceivable. We must admit fallibility; not every situation which seems conceivable will turn out, upon further examination and theorizing, to be metaphysically possible. But as long as we are careful not to postulate sloppily what we are imagining — we ought to take it to follow that what we imagine is metaphysically possible. It is a route to metaphysical possibility; it is not simply equivalent to it, that is there may be some metaphysical possibilities that are inconceivable to us. But it is our only route to metaphysical possibility.

The real worry for those who espouse the intuitions, then, is not that they need an argument from conceivability to metaphysical possibility, but whether what they seem to be imagining is genuinely conceivable. It should be clear, then, that I

want to emphasize that the point is not that whatever we think is conceivable is thereby possible. There is no hard and fast rule for deciding which apparently conceivable situations are genuinely conceivable. However, if we are careful in describing what we think we are conceiving, we should and do take this to be some reason to think that the relevant situation is metaphysically possible. We should and do take (carefully described) intuitions of conceivable distinctness to be counterexamples against the requisite metaphysical identity, admitting fallibility all the while. If this is the case, then to claim that nothing metaphysical follows from epistemological possibility — as if we never need take seriously any argument from conceivability — is to sidestep a real challenge. We should rather grant that it is a metaphysical challenge, and answer it as such.

Perhaps I have been belaboring this point, but it seems that the opposite point is one that is often accepted, even welcomed, and not clearly argued for.[16] Levine points out that since the intuitions Kripke points to are 'merely an epistemological matter', we should thereby *expect* them to be devoid of metaphysical implications. Proponents of the point seem unaware that they are committing themselves to the claim that we cannot draw, legitimately, any conclusions from our concepts. The cautious part of my point here is that this is a claim that needs to be argued for. The stronger claim is that it is not clear it can be plausibly argued for, given that its proponent needs to tell us what would have metaphysical implications if not epistemological entities such as concepts, and the intuitions that ensue. One way to argue for the point is to argue for Cartesian skepticism; and this would surely disallow the proponent from putting forward any sort of claim such as 'physicalism is most likely true'.

Suppose Levine tried to object that again I am misinterpreting him, and what he means by 'epistemic possibility' is something like 'merely apparent epistemic possibility'. In other words, he does not think that nothing whatsoever follows from epistemic possibility: what follows is something broader than metaphysical possibility,

call it *imaginable possibility*. Some situations may be imaginably possible, even though they are impossible; such is the situation with the intuitions of contingency and physicalism. But since what is imaginably possible is not thereby genuinely possible, we need not concern ourselves about it when we are devising theories. Theories may be true, though they need not dispel intuitions imaginably possible impossibilities. My metaphysical possibility, he may say, is just this broad imaginable possibility, and as such is no concern to those trying to theorize about the mind. There are two things to be said in reply. First, what I mean by metaphysical possibility is not imaginable possibility. It is true that some impossible situations may seem possible to us, if we are radically confused or ignorant of certain facts. But there is nothing broader than metaphysical possibility. Second, suppose we grant that what is only seemingly conceivable, does only imply "imaginable possibility", where this may be some metaphysical impossibility. I would respond as I did above to the weakened version of L3*. Either such intuitions point to the falsity of physicalism, or they do not. If such intuitions of imaginably possible impossibilities were never to subside, what reason would we have for assuming that that they did not actually point to the falsity of physicalism? On the other hand, if they do not point to the falsity of physicalism, then why should it concern us if some of us claim to be able to imagine impossibilities? Given that they imply contradictions, why should we concern ourselves with such imaginable impossibilities when theorizing about the world, especially if they are not a threat to physicalism? It seems it would be better to concern ourselves with distinguishing such impossibilities from the genuine possibilities, that is, it would be better to try to reach a conclusion about what is genuinely possible. If what we are dealing with is an ignorant person like myself who claims to be able to imagine, in a way, that water is not H_2O, we need not take such a person seriously in constructing our theory. Then presumably we need not take an imaginable possibility to be a

substantial enough worry to support something as allegedly problematic as the explanatory gap. So this sort of objection will not take Levine very far.

Someone might object that my own criticism of Kripke's argument concerning the intuitions seemed to rely on the point that those intuitions lack metaphysical implications. After all, I replied to the argument that what leads to the intuitions may just be that we do not *know* enough about which mental states are which brain states. But, so the objection goes, doesn't this just amount to my saying that those intuitions point towards merely an epistemological possibility? I disagree. My point is not that the original intuitions lack the metaphysical implication that mental properties are distinct from physical properties of the brain. My point is rather that at this stage of theorizing, it is up for grabs whether those intuitions are of a genuinely conceivable situation, or a merely apparently conceivable one. Either way, on my view, the intuitions have implications. If the intuitions are of something that is only apparently conceivable, then one is not imagining what she thought she was; but nevertheless what she seemed to imagine had metaphysical implications, that were necessarily false. That the implications are necessarily false will, if they are thus, be borne out by further theorizing.

So, someone might go on to object, I seem to be claiming that we can imagine impossibilities. For suppose it is impossible that pain is not c-fiber stimulation. Then what exactly are Kripke , Levine and others imagining? We all agree we can have intuitions of falsehoods; but how can we imagine what is necessarily false? There is no coherent description of someone who believes the impossible. This is admittedly a deep problem concerning the limits of belief; it is moreover a problem which pervades our theorizing about the world. As such it does not directly challenge my point that what is genuinely conceivable does have implications, and so the real worry for those who espouse the intuitions is whether they are genuinely or merely apparently conceivable. However, I further wanted to say that what is merely apparently conceivable as well

has implications, which are necessarily false. So something more needs to be said. I am not saying that we can imagine impossibilities, but rather that impossibilities can seem coherent.[17] It is uncontroversial that we can be and often are unaware of the implications of our beliefs. The same point applies here. The point is not that we can knowingly imagine the impossible, but that what is impossible can seem possible to us; and its apparent possibility has (false) metaphysical implications. We can have intuitions with metaphysical implications, i.e. something might seem possible, while those metaphysical implications are, unbeknownst to us, (necessarily) false. That they do have such implications is shown by our losing the intuitions once we see that the implications are false. Take for example those who used to think that there were two distinct stars, named 'Hesperus' and 'Phosphorus'. They certainly thought that they were imagining that Hesperus and Phosphorus were distinct. They seemed by all accounts to be imagining it; but they were wrong, since it is impossible for Hesperus not to be Phosphorus. They seemed to be imagining something, which was unbeknownst to them necessarily false, and their intuitions that there were two planets had the following metaphysical implications: that there were two planets, and that Hesperus was not Phosphorus. They were unaware of the contradictory meaning of this last clause, but it was there just the same. Again, seeming to imagine what is actually impossible does have metaphysical implications. I cannot give a general account of how it is that we manage to have false intuitions of merely apparent contingency. The way in which we do so will depend on the particulars of the context.

The present burden is then to explain how the implications might be necessarily false, in the present context of pain and c-fiber stimulation; and this is what I cannot explain in great detail, given the lack of a theory. (This, it should be noted, is no cause for alarm. In the Hesperus/Phosphorus case, and the heat/mke case, we have well-established theories to rely on with which to state our case. This is, as we all know, not so for the mind.) All that can be said at this stage is that once we have

identifications mapped out between many sorts of phenomenal properties and physical properties of the brain, and there will (if the identity thesis is true) be distinctions in the physical properties that will allegedly account for the phenomenal differences. Then it would seem reasonable to identify the phenomenal property with the physical property, especially if the theory were fleshed out in detail, and we have fuller knowledge of what it was about a particular property of the brain that was being claimed responsible for its experiential nature. Perhaps, even then, some will be able to imagine that a world of ghosts and nothing else (ignoring for the moment that this might be impossible as well); but the way to explain this would be to say that the possibility that there are ghosts is not the possibility that we might have been ghosts. The ghosts and their mental properties are not ultimately relevantly similar to ourselves and our mental properties. We may think we can imagine what it would be like to be such ghosts; but it seems that it would be rather difficult to spell out this experience in much detail, without relying on the experiences given through our (bodily) senses. We experience through our nerves, and our brains; what would such ghosts experience their world through? These points may seem silly but their purpose is serious: it really is no easy matter to simply say: imagine we had all the evidence we could possibly have, and even then the mental would seem distinct from the physical. For there is a lot of work to be done in garnering 'all the evidence we could possibly have'.

4. Conclusion

I have looked at two arguments for the explanatory gap, i.e. for the view that physicalism, while probably true, nevertheless does not satisfactorily explain the nature of mental states, in that it does not make intelligible to us why a particular sort of mental state feels as it does. I have argued that neither argument succeeds. McGinn's fails because it relies on a strong and implausible premise that there are only two ways we could come to fully understand the answer to the mind-body problem, by introspecting consciousness, or by studying neurophysiological data alone. This

premise may seem plausible to McGinn if he is conflating it with a distinct claim, that if physicalism is true, then given the neurophysiological facts, the facts about consciousness must be determined. This latter claim, while true, nevertheless does not entail that the only way to discover the link between the mind and the brain is by ignoring either mental or physical data.

Levine's argument ultimately fails for two reasons: first, he relies on the implications of intuitions that arise too early in the process of theory construction; and second, he fails to appreciate that claims about what is conceivable do have metaphysical implications. Levine claims that no matter how rich the neurophysiological story, intuitions that the mental state is conceivably distinct from the brain state would remain steadfast. I claim that we do not have enough of a theory at this stage to be able to make such a statement; in fact it seems more likely that if the theory were fully spelled out, someone who espoused such intuitions would be unable to support the intuitions in any substantial way. Furthermore, Levine's attempt to distinguish the explanatory gap from a metaphysical challenge to physicalism does not succeed. In attempting to argue that the intuitions have no metaphysical implications, he (unwittingly) arrives at an implausibly strong conclusion, i.e. that none of our concepts have metaphysical implications. This sort of broad skepticism, however, while it is a philosophical problem, is not a problem that is in any way specifically worrisome for philosophers of mind concerned with the mind body problem. Barring skepticism, the intuitions he relies on, and which he claims lack metaphysical import, in fact do have the intended metaphysical implications (whether these implications are true or not) and should be treated as such by the physicalist. As such, the explanatory gap is no further concern for physicalists, over and above the original argument given by Kripke.

Chapter 3

A Defense of Property Dualism:

Exploring the Implications of Intuitions of Contingency

0 Introduction

Are mental properties[18] identical to physical properties? Many of us have intuitions that, for example, my having a headache is not identical to the firing of a particular group of neurons in my brain. The correlation between these properties seems contingent, since it seems metaphysically possible to have a headache without having any neurons firing, and conversely. These intuitions of contingency are sometimes exploited by philosophers in various ways, to argue against the materialist thesis that mental properties are identical to neurological properties of the brain, or to argue against the explanatory power of the such an identity thesis.

It is arguable that such intuitions about the distinction between the mental and the physical, while their content is such that they are at odds with the materialist identity thesis, nevertheless do not constitute decisive refutations of that thesis, given our legitimate methods of conceptual revision during theory construction. In other words, one way of responding to these intuitions is just to say that they should and will evaporate, once we have a more fully formed theory of precisely which mental entities (presently conceived of in phenomenal terms) are identical to which physical properties (presently conceived of in non-phenomenal terms). The feeling of contingency could well be a result of our present state of relative ignorance, rather than a result of our intuiting a distinction between mental and physical properties. However, such a reply itself is no decisive refutation of the suggestions of the intuitions. The reply is rather a promissory note, which asks us to wait and see what further theorizing turns up. Given the possibility of offering such a promissory note, the strength of the intuitions on their own is decisively weakened; but nevertheless we should admit that there may be some truth to those intuitions, which may itself be borne out in the course of further theorizing.

A further, but related question is whether mental properties are in some sense reducible to physical properties. The answer to this depends on what it is to be reducible in the intended sense. If by 'reducible' we mean identical to, then clearly the intuitions speak against the reducibility thesis insofar as they speak against the identity thesis. However, there may be other sorts of reducibility that are of interest: for example, the reducibility of the causal powers of one sort of entity to the causal powers of some other entity; or, the sense in which a chair is reducible to (though not strictly identical to) the stuff that composes it. How we conceive of the aforementioned intuitions will in turn affect what we say about whether there is any interesting sense in which the mental is reducible to the physical.

A final question is whether there is any sort of middle road, that somehow manages to satisfy both materialists, who think that ultimately the world is physical, and those who think there is something to the intuitions of contingency. This is an important question to try to answer, I think, since an affirmative response would amount to a dissolution of the mind-body problem, or at least to a promise for future dissolution given further theorizing.

In this paper, then, I intend to explore what sort of a theory could result were we to take the intuitions of contingency at face value, and try to develop a sort of property dualism, according to which mental properties in some way supervene on certain physical properties of the brain, but are nevertheless not strictly identical to those states. Crucial questions I wish to address include whether mental properties are irreducible in any interesting sense; what property dualism amounts to; what is involved in taking the intuitions seriously; and whether being a property dualist of a certain sort would satisfy someone who takes the intuitions seriously. First, I will present Kripke's argument that the physicalist's identity thesis seems contingent (and therefore seems false). Then, I will further examine the intuitions of contingency upon which Kripke relies, and delineate five different sorts of intuitions that we might have

in thinking that the identity thesis seems contingent. I will then discuss the various replies that a materialist could offer to these various sorts of intuitions, and claim that some of the intuitions may not be offensive to the materialist, while will necessarily be rejected. At this point, I will go on to examine John Searle's view of biological naturalism, as an example of the sort of view of someone who claims to take the intuitions seriously, and who claims to provide an acceptable middle road view that should appeal to both materialists and supporters of the intuitions. I will argue, however, that it is obscure what type of intuition, given my classification, Searle takes himself to be relying on: none seems to be able to justify both of his crucial claims, that the physical causes the mental, and that mental properties are harmlessly emergent in the same way that properties such as solidity and liquidity are emergent. I then go on to attempt to delineate my proposed view, that the property of being in pain is the property of being composed of certain physical entities that instantiate certain physical properties, and likewise for other mental properties. My thesis will be that this view is a sort of property dualism, distinct in important ways from Searle's view, that should satisfy many materialists as well as some who find themselves struck by the intuitions of contingency. I begin with a review of the intuitions themselves, and of Kripke's argument concerning them.

1 Kripke's argument against the physicalist

Kripke argues, in the third lecture of *Naming and Necessity*, that given that all identity statements between rigid designators are necessary, the identity statement proposed by materialists must be necessary. However, he argues, there is a clear feeling of contingency to a statement expressed by sentences like 'pain is c-fiber stimulation'. We may grant that pain and c-fiber stimulation are somehow correlated; nevertheless, this correlation seems contingent, as it seems imaginable that the correlation fails to hold. Further, if it is genuinely imaginable that the correlation fails to hold, then the correlation cannot be the relation of identity. However, we must ask, what is involved

in imagining that the correlation fails to hold? How we reply to the intuitions, and whether we finds them plausible, will depend on how the intuitions themselves are cashed out. Kripke cashes them out in the following way: The statement that pain is c-fiber stimulation is contingent, because it is seemingly coherent that one should experience pain, without undergoing the firing of c-fibers, and vice versa (Kripke 1972 , pp. 146-7). Kripke challenges that

> Someone who wishes to maintain an identity thesis cannot simply accept the Cartesian intuitions that A can exist without B, that B can exist without A, that the correlative presence of anything with mental properties is merely contingent to B, and that the correlative presence of any specific physical properties is merely contingent to A. He must explain these intuitions away, showing how they are illusory. (1972 p. 148)

However, Kripke maintains, explaining the intuitions away is not easy. He offers us a paradigmatic case of explaining away similar intuitions, but notes that this will not help in the case of the intuitions with which we are presently concerned. The similar intuitions involve the alleged reduction of heat to molecular kinetic energy (mke). Suppose someone were to claim that she could certainly imagine that heat exists while mke does not; the situation would be one in which people could feel sensations of warmth, but these sensations would be caused by some other phenomenon. Kripke replies that such a situation may seem offhand to be one in which heat exists but mke does not, but in fact it is not such a situation. Rather it is one in which something other than heat causes warm sensations. Furthermore, this should be expected. For given our established discovery that heat is mke, we really cannot coherently imagine one without the other. What causes the confusion is that we mistakenly take heat to be identical to some contingent property by which we identify it, in this case its property of causing in us a certain sort of sensation.

This sort of reasoning is of no avail to those who wish to identify pain and the stimulation of c-fibers, says Kripke. For when we think of a pain, we do not identify it in virtue of its contingent property of producing some sensation in us; the sensation, rather just is pain. The feeling of pain is essential to it, because it is identical to it: the feeling of pain just is pain. So we cannot claim that we are imagining a situation in which we only think we are experiencing pain, but we are not. So, as it stands, the statement that pain is c-fiber stimulation not only seems contingent, but seems to be stubbornly so, rather than necessary, and so taken as a strict identity statement it must be false. The argument stands as a challenge to the materialist to explain how a statement identifying the mental with the physical, a statement that seems to be necessarily false, could possibly be true.

This is roughly the argument, which Searle points to as the 'decisive' refutation of the claim that mental properties are reducible to physical properties (Searle, p. 118). I do not agree that the argument is decisive; there are various responses that could legitimately be made on the part of the physicalist. The reply that I favor, as I mentioned above, is that we may not be able to explain away the intuitions at this stage of theorizing; but this does not mean that in the future the intuitions will continue to stick, once we have each correlation between the phenomenal properties and the relevant neurological properties mapped out. In fact, Kripke's explanation regarding heat seems to rely rather crucially on the fact that given our current theories, we know that heat just is mke. In cases where no such theoretical identity has been established, such intuitions are on shaky ground. However, I won't discuss this point further here. For I do think that the intuitions themselves are fairly striking, and it might be open to Kripke to charge that my exhortation to wait and see does not give the intuitions their due respect. It does seem strange that a feeling, a phenomenal experience, is strictly identical to the firing of neurons. However, and importantly, there may be some other way to make sense of the feeling of contingency, that does not rely on spelling out the

intuitions in the way that Kripke does. To get clearer on this, we need to examine the intuitions more closely, and what might be serving as a basis for them.

So I propose that we take the intuitions at face value and see where they might lead. First, then, we must explore just what the intuitions amount to, and how Kripke himself uses them.

2 Classifying the intuitions of contingency

What exactly does it mean to say that we can imagine that pain and c-fiber stimulation exist without each other, and how exactly does the feeling of contingency arise? Kripke's exact words when he brings up the intuitions against the type identity theory are as follows: '...the identity theorist is committed to the view that there 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...' (Kripke 1972, p. 149). This, however, does not illuminate the intuitions or elaborate on their grounds. His earlier remarks regarding the intuitions as against the token identity thesis are likewise fairly minimal: 'it is at least logically possible that B [C-fiber stimulation] should have existed ... without Jones feeling any pain at all' (Kripke, p. 146). Again the intuitions may seem plausible, but there is more than one situation that one might be imagining, in envisioning pain without c-fiber stimulation. The only suggestive thing he seems to say about the nature of the intuition that c-fiber stimulation could exist without pain, is to distinguish it from the intuition that pain could exist with c-fiber stimulation. This latter he claims to be closer to the 'Cartesian consideration' (Kripke 1972, p. 147). This distinction between the two sorts of intuitions may seem curious, depending on what the basis is for having the intuitions in the first place. It would be helpful, then, to try to sort out the various sorts of intuitions we might have, in thinking that the relation between the mind and brain is contingent. I think there are roughly five sorts. (The last two sorts of intuitions may be more

appropriately construed as responses to the first three sorts of intuitions put forward by an anti-physicalist.)

Type 1 Intuitions. Suppose what we are imagining is the following: mental properties exist, but no physical properties or entities. In other words, there are only subjective, conscious experiences of pain, pleasure, despair, euphoria, and everything in between. Let's call these sorts of intuitions 'pure Cartesian intuitions', for what we envision is the world of Descartes' skeptic: it seems as if there is an external world, but all that exists are experiences. This is the Cartesian notion that there could have been feeling, thinking, experiencing souls, even though no matter existed whatsoever. Note that if someone has intuitions of this sort, then pretty clearly he will not find the identity thesis plausible: pain is not identical to c-fiber stimulation, if nothing remotely like c-fiber stimulation need exist for pain to exist. Whatever pain is, according to these intuitions, we can isolate it from any sort of matter, and it furthermore it can exist independently of matter. Moreover, on such a view, it is straightforward where the contingency lies. The statement that pain is correlated with c-fiber stimulation is contingent, on such a view, because it is only contingently true that the two sorts of things are in fact correlated in this world. So, on this view, the statement that pain is c-fiber stimulation is only true if taken not as a statement of strict identity, but as a statement of some sort of predication: in this world, mental properties happen to be physical properties, in some way or other, in virtue of instantiating some (non-essential) physical property. These Cartesian intuitions, moreover, suggest a substance dualist view: they seem to rely on a thought that matter and mental properties are radically distinct in nature, so as to be different substances altogether. For if mental properties can exist with no physical properties or entities, there must be mental stuff for them to be instantiated in.

Type 2 Intuitions. Contrast the above with the following intuitions. Suppose we find the above situation close to nonsense; how is pain supposed to exist,

while it has nothing, no matter, to exist in? Perhaps we think there must exist some matter or other, for there to be experiences: mental properties must co-exist with physical objects, but the correlation between them can vary widely. So suppose we ground our intuitions of contingency in a hypothetical situation like the following: I wake up with a terrible headache, but when the doctor reviews the cat scan, it turns out that my brain is made of wood. (Further, this is no surprise in this world, and all brains are made of wood.) Call these sorts of intuitions 'mixed Cartesian intuitions'. Once again, the contingency arises in that there is no necessary connection between the existence of pain and the existence of any particular sort of matter (since there is nothing special about wood, and one who has such intuitions would most likely readily substitute any other sort of matter in its place). These intuitions are distinct from the first intuitions, in that according to the first intuitions there needn't be any matter at all, while according to these second intuitions, matter must exist in order for the mental properties to be instantiated. Once again, pain turns in out in this world to instantiate some non-essential physical property, of involving the firing of neurons, and so on, but it could have arisen out of mere wood. Again these intuitions support either a property dualist view, or even a substance dualist view; the view they support is not very interestingly different from the view above. For if we suppose that mental properties have no necessary connection whatsoever to the physical objects in which they are instantiated, it is a small step, I take it, to take to suppose that mental stuff exists in such a world as well.

Type 3 Intuitions. The complement to the above two views is that which envisions a world of matter and physical properties, but no experiences. Moreover, what is allegedly imagined is a spooky world of complicated matter, matter organized the very same way that our matter is organized, and still -- no conscious thoughts, no feelings, no experiences. These we may call 'zombie intuitions'. Such intuitions again support substance dualism, or at least property dualism, and are against functionalism,

the view that the mental is characterizable solely in terms of its causal and relational properties. The contingency arises on such a view in that again there is merely a contingent correlation between mind and matter: there is no necessary connection at all between mind as we know it and matter as we know it, according to . For given our world as it is with its matter organized as it is, there is some twin world in which there are no mental experiences going on. As such these intuitions may go hand in hand with the first two sorts of intuitions.

Type 4 Intuitions. Putting aside for the moment the credibility of the above three sorts of intuitions, let's contrast them with allegedly less radical intuitions. Suppose we find instead the following sort of situation imaginable: human beings do not exist, and moreover no c-fibers exist. Instead there is a race of Martians ruling the galaxy, and they've got silicon-based d-fibers. Or, similarly, there are robots with brains not of flesh and blood but made of computer chips. In either case, there is sentience, although it arises in a very different way than it does in our world. While, against the type 2 intuitions, a world made of blocks of wood lying around randomly would not be a world with pain in it, a world where there were blocks of wood organized in the right way, with enough complicated connections, would be a world with pain (perhaps pleasure as well). Call these intuitions, which support functionalism, 'multiple-realizability intuitions'. This construal of the intuitions of contingency can be seen as a sort of response to the first three intuitions, in that it explains the feeling of contingency in a way which does not on its face commit one to property or substance dualism: the claim is that while we might think the relation between the mental and physical is contingent, nevertheless this feeling may be arising not out of the plausibility of any sort of dualism, but out of the computational nature of the mental. The computational nature of the mental is not at odds with its material nature; it amounts to the ability of computational states to be grounded in or run on various sorts of 'hardware'.

Type 5 Intuitions. Are there other ways to make sense of the intuitions that Kripke alludes to? I think there are. Call these 'Lockean intuitions'. Recall Locke's point that a tree is not identical to the molecules which make it up (in his Essay Concerning Human Understanding). This is so because we can remove some of the molecules, while the tree itself remains. Trees, then, are *composed* of matter in a certain organization; so we might conjecture that the property of being a tree is the property of being composed of various sorts of matter which instantiate various physical and biological properties. On this last sort of construal, the analogous intuition that pain is distinct from c-fiber stimulation is akin to this Lockean point. Suppose that what we imagine when we experience the intuitions is not that pain could exist without matter, or that pain could exist without our sort of matter, or that our matter in its organization could exist without pain; rather we imagine that given any particular pain, say the headache I woke up with this morning, it is not strictly identical to the stimulation of the c-fibers, because we could have had the exact same pain even though one or two fewer c-fibers were stimulated. On such a view, we grant with the materialists that pain and c-fibers (or whatever the appropriate neurological property turns out to be) are very intimately related. In fact, the relation between pain and c-fibers is not a contingent one at all; it is that of supervenience. Given that there is pain, there must be c-fiber stimulation, and given (a certain amount of) c-fiber stimulation, there must be pain. What provides the contingency is rather the contingency of the statement that a particular instance of pain is composed of the stimulation of c-fibers 1 through n. This statement is contingent because the pain could have remained, while c-fibers n through n- 5 failed to be stimulated.

What of pain, itself, regardless of some specific instance of it? On this contrual, this is simply the property of being composed of the stimulation of c-fibers, in some certain organization or pattern. (Presumably not just any stimulation of c-fibers in

isolation would compose a pain.) So, again, pain is not strictly identical to the stimulation of c-fibers; it is the sort of thing the instances of which are composed of the stimulation of c-fibers. As such, I take this to be a property dualist sort of view, in that it distinguishes the mental property, pain, from the "physical" property, the stimulation of c-fibers. But, as we can say that trees are nothing over and above the hunks of matter in a certain organization that make them up over an extended period of time, we can also say that in this sense, pains are nothing over and above the stimulation of c-fibers that compose them. Thus the view is also quite acceptable to the materialists, in that it does not posit a profound distinction between mental stuff and physical stuff, or between mental and physical properties. It is property dualist in a relatively minimalist way, in that mental properties are higher-order properties that are distinct from lower-level neurological properties of the brain. Nevertheless, just as we can say many interesting things about trees without having to mention their composition, so we will most likely be able to say many interesting things about pains without mentioning their composition. This does not mean, however, that it is not essential to pains that they have the composition that they do; it only means that pains have many other properties as well.

Note that this is a very different way of cashing out the intuitions. Again, this construal of the feelings of contingency can be seen as a sort of response to the first three construals: we sense an air of contingency, and therefore an air of falsity, to the identification of pain with c-fiber stimulation, while those intuitions nevertheless do not arise as a result of the plausibility of (irreducible) property or substance dualism. Someone could have intuitions along the lines of intuition 5, even if he found the other construals of the intuitions highly implausible, or thought them the sort of intuition that should fade once further theorizing takes place. While the first three sorts of intuitions seemed to rely on a radical distinction between matter and mental properties, these last intuitions rely on no such dichotomy. If pains are in some way made up or composed

of neurons firing, they cannot be in any clear sense radically distinct from neurons firing.

Which sort of intuitions does Kripke have in mind? Clearly, he has in mind one of the first three, and does not have in mind the Lockean or the multiple-realizability intuitions. He writes in a footnote on page 145 that

> it can be argued that a statue is not the hunk of matter of which it is composed. In the latter case, however, one might say instead that the former is 'nothing over and above' the latter; and the same device might be tried for the relation of the person and the body. The difficulties in the text would not then arise in the same form, but analogous difficulties would appear. [Such] a theory ... would have to hold that (necessarily) a person exists if and only if his body exists and has a certain additional physical organization. Such a thesis would be subject to modal difficulties similar to those besetting the ordinary identity thesis...

In other words, if we try to redescribe the intuitions of contingency by claiming that mental properties are properties of being composed of physical objects that instantiate certain physical properties, this identification will again seem contingent, and therefore false as a necessary identity. Presumably, what he means here is that we could always imagine that minds could exist even though they were composed out of very different materials. This sort of objection will have to put off until later on in the paper, where I will try to deal with objections; I bring it up here to emphasize which sorts of intuitions Kripke has in mind. The response of a materialist who found intuition 5 plausible would be that such intuitions will (hopefully) subside as more theorizing is brought to bear on the matter.

In fact, such a response could be given, by a materialist who finds the identity thesis plausible, to the first three sorts of intuitions. At this, point, then, before we move onto Searle's view, and which sorts of intuitions he might have in mind, we should note that a materialist's response to the intuitions will vary greatly depending on

which intuitions are being espoused. To the first three, he might feasibly reply that these intuitions merely reflect our present state of ignorance, and that they should and will evaporate, at least among those in the know, once we fully develop our theory of the mind.[19] To the multiple-realizability sort of intuition, a physicalist could also claim that given intuitions like those espoused by Ned Block[20], that if the nation of China were somehow arranged so as to instantiate the same functional property that a subject of mental experience does, nevertheless that nation of China would not be a conscious entity, that the functionalist thesis seems implausible. However, this would not be a decisive refutation of the view, particularly since the functionalist could retort that perhaps given further theorizing such intuitions would evaporate. To support his point, then, a materialist may argue that it is unclear why we should expect these intuitions to subside, given that it is unclear why we should ignore the importance of the particular neurophysiology of the mind in this instance of theorizing, while in any other sorts of theorizing we do look to the biological structure of the entity in question, to learn more of its essential nature. (This is not intended as any sort of refutation, but I will not go further into this debate. I take this debate to remain open, though I believe there is more evidence supporting the biological side.) In fact, while the functionalist view that the multiple-realizability intuitions support is not substance or property dualism, in some ways it is arguably not very far off. For where does the contingency arise? It arises again the relation between the existence of pain and the existence of (not just stimulated c-fibers but) any kind of matter. The caveat is of course that the matter must be organized in the right way; but again it is unclear why the organization, rather than the matter itself, should be taken to be the determining factor.[21]

To the Lockean intuitions, however, it may be that the materialist would welcome them as a reconstrual of the Kripkean intuitions of contingency, for even though they may support a sort of property dualism, it is not a property dualism according to which mental properties are profoundly different from other sorts of

physical properties; it is a sort of property dualism which should be amenable to materialists. This last claim is one which John Searle takes himself to be able to make: he claims that his view is not really property dualist, since he wants to say that mental properties are as harmless as other sorts of higher-level physical properties. At this point, I will present his view, and then go on to argue that the view is unstable given internal tensions, and that he cannot in fact legitimately claim that his view should be so welcomed by physicalists.

3. Searle's biological naturalism: an unstable hybrid view

3.1 Statement of Searle's View

Searle's statement of his view of conscious mental properties is as follows:

> The brain causes certain "mental" phenomena, such as conscious mental states, and these conscious states are simply higher-level features of the brain. Consciousness is a higher-level or emergent property of the brain in the utterly harmless sense ... in which solidity is a higher-level emergent property of H_2O molecules when they are in a lattice structure (ice), and liquidity is similarly a higher-level emergent property of H_2O molecules when they are, roughly speaking, rolling around on each other (water). (p. 14)

From the above, we can isolate two main claims. One is that the brain causes conscious mental phenomena, and the other is that conscious mental phenomena are higher-level in the same way that phenomena such as liquidity are higher-level. I do not think that Searle can easily embrace both of these claims. For, as Paul Churchland notes in his review of Searle's book[22], liquidity is not in a strict sense caused by the motion of water molecules; rather, it is constituted by it. I would add to this point that whether we think liquidity is simply identical to the motion of molecules, or rather that it is the property of being composed of molecules that are instantiating certain physical properties, in either case liquidity is not caused by the underlying physical entities and their properties. For this reason, if mental phenomena are caused by the brain, they cannot

be straightforwardly the same sort of properties as other higher-level phenomena. Nevertheless, if we take constitution, or composition, to be distinct from strict identity, as I think we should, mental phenomena and phenomena such as liquidity will be similar in the following way: neither is strictly identical to the requisite underlying properties. However, if we do take them to be similar in this way, as Searle wishes to, the reasons for the distinction will be similar as well. In other words, mental phenomena cannot be both caused and composed by the underlying properties. One or the other but not both must be the reason for the distinction between mental phenomena and the neurological properties that underlie them. This tension becomes clearer as we consider how it is that Searle takes himself to be relying on the intuitions of contingency, and exactly how he might be taking these intuitions to support his claim that mental properties are irreducible.

3.2 Searle on the irreducibility of mental properties

3.21 Two sorts of reducibility: causal and ontological

Searle tries to make two main points about conscious mental properties and reducibility; the first is that mental properties are not ontologically reducible, i.e. not identical, to neurological properties of the brain; the second is that this irreducibility 'has no deep consequences' for our view of the world. In this section and 3.11, I will examine the first of these claims, and show that given Searle's commitment to causal reduction, it is unclear how he can avoid a commitment to some sort of ontological reduction. In section 3.12, I will challenge his claim that the irreducibility has no deep consequences, in part by expressing puzzlement over what he means by this; and in section 4 I will try to show how the property dualism grounded in Lockean, rather than Cartesian, intuitions can overcome these difficulties. First, we shall deal with Searle's claim about the irreducibility of mental properties.

Searle distinguishes several types of reduction, two of which will concern us here. He is concerned to show that while he holds that mental properties are

causally reducible to physical properties, they are nevertheless not ontologically reducible. I want to show that while it is clear why he wants this to be the case for his view, it is not clear that he can ultimately embrace both claims. First, what does Searle mean by reducibility? He defines the relevant different sorts as follows (Searle, pp. 113-114):

> According to Property Ontological Reduction (POR), properties of one type are found to consist in properties of other types. (An example of this is that of heat being reducible to the mean kinetic energy of molecular motion.)
>
> According to Causal Reduction (CR), two types of things (or properties) are related such that the causal powers of the reduced entity are found to be explainable completely by reference to the causal powers of the reducing entity.

Searle is concerned to show that his commitment to CR does not commit him to POR, since as he sees it a commitment to POR would amount to a commitment to materialism. First, however, there is the question of what exactly it is that CR says, and in particular what Searle means here by 'explainable'. An initially plausible interpretation is the following: one type of thing or property is causally reducible to another, just in case the causal powers of the former are explainable completely by reference to the causal powers of the latter, because the causal powers of the two are simply identical. This seems to be on its face what Searle has in mind, since he does seem to want the causal powers of the mental and the physical to be one and the same: he claims that he does not have in mind a sort of view of mental properties according to which 'once [consciousness] has been squirted out [by the brain], it then has a life of its own' (Searle, p. 112). Unfortunately, however, this is not obviously what he should have in mind. For suppose that the causal powers of pain and c-fiber stimulation are simply identical. Then this would seem to be some reason to believe that pain and c-fiber stimulation themselves are identical, and this would mean a commitment to POR, which Searle clearly wants to avoid. Moreover, how can he avoid this sort of

identification of mental and physical properties, if he identifies the causal powers of each? Typically, in fact, a reason for identifying theoretical phenomena is simply that the phenomena share all causal powers. Consider heat and mke: we discover that heat has certain causal powers, for example it causes certain sensations in us, it melts wax, and so on. We go on to isolate what physical entity is in fact the source of such causal powers, and we find that it is mke. Consequently we identify heat with mke. But this is clearly a case of an ontological reduction. Therefore, this straightforward interpretation cannot be what Searle has in mind.

Perhaps, then, he has in mind something like the following: one sort of thing (or property) is causally reducible to another sort of thing (or property) just in case the causal powers of the former are causally explainable fully in terms of the causal powers of the latter, given that the latter causes the former. This is clearly a much weaker principle of causal reduction. That Searle has this in mind is suggested by the fact that the only thing he seems to offer in his support of CR is the claim that on his view, mental properties are caused by neurological properties: 'I hold a view of mind/brain relations that is a form of causal reduction, as I have defined the notion: Mental features are caused by neurobiological processes' (Searle, p. 115). So he seems to take it as a justification of his commitment to CR, that he is committed to a principle of transitivity of causation: if event a causes event b, then whatever causal powers b has are caused by the causal powers of a. And it is true that we might say that, in a sense, whenever one event causes another, the causal powers of the former fully explain the causal powers of the latter, in that they have been caused by them. But first of all, this does not seem to mix well with Searle's claim that consciousness has no life of its own; and further, it is too weak a notion to be very helpful in constructing an explanation of the mental. For while it may be that when one event causes another, the causal powers of the latter are caused by the former, in general there will be causal powers of the latter that are not explained, in an interesting sense, by those of the former. Surely there are

cases in which even though one event causes another, there are causal powers of the second event that are not in an interesting sense fully explainable in terms of the first event. If I get into an accident because my tire blows out, it seems unlikely that any causal power of that accident will be explainable in terms of my forgetting to check the air pressure, though it will be true that in a sense it will have been caused by that event. In other words, on this weaker interpretation of CR, events will more likely take on causal significance of their own, and as such won't be fully explained by the events that caused them. Again, there seems to be a tension between Searle's holding CR, and his denial of POR. This tension is exacerbated further by his slippery use of the Kripkean intuitions of contingency. We have noted already, with Paul Churchland, that there is tension in claiming both that mental properties are caused by physical properties and that they are relevantly similar to other higher-level properties like liquidity. A further point is that the argument resting on the intuitions of contingency, which Searle uses to support his view that mental properties are distinct from physical properties, cannot support all aspects of his view. As we shall see in the next section, Searle cannot help himself to the intuitions of contingency while at the same time hold that the causal powers of mental properties are explainable, in an interesting sense, in terms of the causal powers of physical properties.

3.22 Can Searle help himself to the intuitions of contingency?

Why is it that Searle takes himself to be committed to a denial of POR? He thinks that the intuitions of contingency as discussed by Kripke and others show decisively that POR is not true of mental properties (Searle pp. 117-118). We have seen that there is tension between Searle's claim that CR holds while POR does not; for given our discovery that, say, the causal powers of heat could be entirely explained by the causal powers of molecular kinetic energy was best attributable to the fact that heat just is molecular kinetic energy, either the same would seem to go for heat and c-fiber stimulation, or the causal reduction would not seem to hold. Given this tension, the

question that springs to mind at this point is how is it that Searle is conceiving of the intuitions, as a reason for denying POR? What sort of basis grounds the intuitions that support both the claim that mental phenomena are just like other higher-level phenomena, without conflicting with the claim that mental phenomena are caused by and therefore are distinct from underlying properties of the brain? Searle himself does not spend too much time examining the intuitions themselves, and what he takes himself to be imagining. He sums up the anti-reductionist argument fairly crudely, asking us to 'suppose we tried to say the pain is really "nothing but" the patterns of neuron firings. Well, if we tried such an ontological reduction, the essential features of the pain would be left out. No description of the third-person, objective, physiological facts would convey the subjective, first-person character of the pain, simply because the first-person features are different from the third-person features' (Searle, p. 117). As far as I can see, more time needs to spent considering how he could be using the intuitions to support this difference between 'first-person' and 'third-person' features, and whether this use will be in conflict with any other crucial part of his view.

Consider, then, the type 1 intuitions of the pure Cartesian. Could these intuitions be of help to Searle? Recall that these are intuitions of its being conceivable and thereby metaphysically possible that mental phenomena exist regardless of whether matter exists. These intuitions clearly support Searle's conclusion that mental phenomena are distinct from material phenomena. But do these intuitions support his conclusion that mental phenomena are higher-level, or emergent, in the same way that liquidity is higher-level? In fact, they do not. For the two sorts of phenomena to be similar, we would as well have to be able to imagine that liquidity could exist regardless of the existence of anything physical. While we can imagine that liquidity exists even though there is no water, we cannot imagine liquidity existing even though there is no physical stuff whatsoever. Were such intuitions plausible, they would indeed support the conclusion that liquidity is distinct from the underlying property of molecules. But

given that our established theory entails that liquidity is a property of physical stuff, we certainly cannot coherently describe a situation in which liquidity exists, but physical stuff does not. So pure Cartesian intuitions cannot support the conclusion that Searle wants, i.e. that mental phenomena are both caused by physical phenomena, and are higher-level in just the same way as liquidity is higher-level.

Moreover, it may seem on its face that the intuitions could form part of a theory in which in this world, mental properties are caused by physical properties, while in other worlds, they are caused by something else altogether (presumably, by other mental stuff and properties); this is what secures Searle the contingency of the relation between the mind and the body. However, on further examination, this is even more problematic for Searle. For if what is grounding the view is something like the pure (or mixed) Cartesian intuition, it becomes fairly mysterious why we should think that all the causal powers of mental properties are fully explainable by reference to the causal powers of underlying properties. For what are we to say, exactly, of the situation we imagine in which my mental properties remain the same, and so presumably the causal powers that these involve, and yet my brain is made of very different material, with very different causal powers, perhaps with no causal powers? Take either reading of the above CR. If what he means is that the causal powers of the mental just are the causal powers of the physical (which is supported by his claim that he does not have in mind a sort of view of mental properties according to which 'once [consciousness] has been squirted out [by the brain], it then has a life of its own' [Searle, p. 112]), then such pure Cartesian intuitions make no sense, if we are imagining that the mental exists just as it does in this world, with its causal powers intact. For if those causal powers exist in the possible world as they do in this world, then they are physical as well; but in the possible world, there are no physical properties or objects. On the other hand, take the weaker interpretation of CR, according to which the causal powers of the mental are caused by the causal powers of the physical. Then, apparently, what

will be causing the mental powers in this other possible world will be something other than the physical causal powers. (And again he will have to face the probability that such causal powers will indeed take on a life of their own.) What are these other properties? Since there is nothing physical in such a world, they must be something non-physical; and again, this would make the mental radically distinct from other 'harmless' higher-level properties such as liquidity. In other words, if we take the Cartesian intuitions seriously, as he seems to want to do, how are we to explain the causal powers of the mental in the hypothetical Cartesian situation? Again, given his commitment to a biological level of explanation, I don't think Searle would want to agree that such a situation (in which my mental properties have causal powers and yet my brain does not) is indeed conceivable; but this merely emphasizes the dubious nature of his reliance on the Cartesian intuitions as put forward by Kripke. Therefore, these sorts of intuitions cannot, appearances to the contrary, support Searle's claim that the causal powers of the mental are fully explained by the causal powers of the physical.

What about the mixed Cartesian intuitions, according to which mental properties must have some physical stuff in which to be grounded, though the physical stuff need not be organized in any particular way, that is, it need not have any physical properties in order to determine the existence of the mental properties? This is a slightly more subtle point to make, but Searle must as well deny the import of these sorts of intuitions. For why is it that creatures like Howdy Doody are not conscious creatures in our world, and what is it about that other world, that makes it the case that Howdy Doody is conscious? One thing he might say is that in that other world, some mental stuff, the soul of Howdy Doody, suffices to determine his consciousness, in which the intuitions collapse into the pure Cartesian intuitions, and again mental properties are relevantly dissimilar to properties such as liquidity. (Moreover, the above concerns about CR reapply to a degree: if the causal powers are identical in our world, then how do they manage to come apart in Howdy Doody's world? If the causal

powers are distinct, Howdy Doody's world would make sense, but again the intuitions would seem to collapse into those of the pure Cartesian.) So it would seem that these mixed Cartesian intuitions are of no help to Searle.

What about the zombie intuitions of type 3? These intuitions may seem of help to Searle in establishing his claim that the physical causes the mental. For while in this world, the instantiation of physical properties are causally sufficient to ensure the instantiation of mental properties, in other worlds, the instantiation of physical properties is not so sufficient. However, there is again the problem that we have no similar intuitions in the case of liquidity. We have no plausible intuition that the world could exist just as it does now, with all its lower-level physical properties in place, while higher-level properties like liquidity are somehow absent. Given what we know about the world, this sort of situation is not coherently describable. Furthermore, however, Searle faces deeper problems, concerning the causal powers of the mental. For if these are simply identical to the causal powers of the physical, then in the zombie world, presumably the causal powers of the mental should exist as well. But ex hypothesi they do not. On the other hand, if the causal powers of the mental are merely caused by the causal powers of the physical, then such a world is possible; but again, it will be difficult for Searle to maintain that the mental does not have a causal life of its own. In effect, appearances to the contrary, Searle must deny the plausibility of intuitions such as these. Intuitions of type 1 - 3 are of no help to Searle, although these seem to be the intuitions that he borrows from Kripke.

What about the multiple-realizability intuitions, that mental phenomena can be grounded in any matter whatsoever, so long as it is organized properly? Given Searle's commitment to the biological level of description, he ought not have these sorts of intuitions in mind. And it should be clear that they could not support all facets of his view. While they could support the claim that mental properties are distinct from neurological properties (since mental properties are on such a view functional

properties), and analogous intuitions could support the claim that the liquidity is distinct from the underlying motion of the molecules, these intuitions cannot support his claim that mental properties are caused by physical properties. For functional properties are not in any clear sense caused by the underlying properties in which they are realized. Again, these intuitions are of no help to Searle.

What about the Lockean intuitions, that a particular instantiation of a mental property is distinct from the underlying neurological property, since it is rather the property of being composed of the underlying physical matter along with its physical properties. Could this intuition provide a basis for Searle's view? While these intuitions could support the claim that mental features are emergent in the same harmless way that features of liquidity are emergent, once again these intuitions do not support the claim that mental properties are caused by the underlying neurological properties. (What they do support, however, is the claim that the causal powers of the mental are explainable in an interesting sense in terms of the causal powers of the physical. On this view, the problem of mental causation becomes on a par with the problem of higher-level causation in general.) To be composed of something is not to be caused by that thing. These intuitions may get Searle some of the conclusions he wants to appease the materialists, but they will not allow him his other claims concerning how mental properties are determined.

The conclusion to draw is that not only is it unclear how Searle means to ground his hybrid view in the intuitions of contingency, it is furthermore impossible for him to succeed (unless he can come up with some other construal of the intuitions). What he seems to have in mind is either the mixed or pure Cartesian intuitions, or those of the zombie; but as we have seen, these cannot secure his view in its entirety, for these would sharply distinguish mental properties from other sorts of higher-level properties such as liquidity. He seems to want the best of both worlds: he wants mental properties to be caused by the underlying properties of the brain, for this would justify the

intuition of contingency we attach to the correlation between the two, and as a result would allow their natures to be radically distinct; and yet he also wants mental properties to be determined by the underlying properties in the same way that the liquidity of water is determined by the underlying properties of water. However, he simply cannot have it both ways.

The problems for Searle's hybrid view, however, do not stop here. What of Searle's claim that the irreducibility of mental properties — the distinction between them and neurological properties — has 'no deep consequences'? What does he mean by such a statement? In his own words, he wants to show 'why it does not make any difference at all to our scientific world view that [mental properties] should be irreducible. It does not force us to property dualism or anything of the sort. It is a trivial consequence of certain more general phenomena' (Searle, p. 116). He seems to want this to be true as a means of making his view palatable to materialists. In this section, I want to argue that it is unclear why Searle thinks his view is not property dualistic; and moreover, it is unclear that the alleged distinction is trivial on Searle's view. It is even less clear why Searle should find this triviality a benefit.

3.23 Searle on the consequences of irreducibility

Searle is concerned to show that his view is not property dualistic, in that he sees mental properties as just the same sort of properties as other higher-order properties like liquidity. We have already seen above that Searle's claim that mental properties are relevantly similar to properties such as liquidity is problematic; however, insofar as that is the case, it will be problematic to deny that his view is property dualistic.[23] He also seems to take it that his view can be distinguished from property dualism if he denies that his view has deep consequences for our world view. The argument Searle offers to show why the irreducibility of mental properties has no deep consequences is really an argument for the irreducibility of mental properties. The strategy is to compare other instances of ontological reduction to the case of the

attempted reduction of mental properties to neurological ones, and conclude that the methodology of the former just does not work for the latter. Therefore, not only are mental properties irreducible to neurological properties, but this fact is a result merely of our methodology of reduction. (The argument is similar to one given by Nagel in Warner & Szubka). The argument is summed up in the following passage:

> Part of the point of the reduction in the case of heat was to distinguish between the subjective appearance on the one hand and the underlying physical reality on the other. Indeed, it is a general feature of such reductions that the phenomenon is defined in terms of the "reality" and not in terms of the "appearance". But we can't make that sort of appearance-reality distinction for consciousness because consciousness consists in the appearances themselves. (pp. 121-122)

The idea seems to be that with earlier reductions, say of heat to molecular kinetic energy, it was crucial to carve off sensations of heat from the phenomenon of heat itself, and then locate heat in the cause of the sensations. Since, however, the argument goes, this practice of distinguishing the appearance from the cause is necessary to reduction, were we to try to perform a reduction of pain, we would be faced with the following, in Searle's words: 'We could simply define ... "pain" as patterns of neuronal activity that cause subjective sensations of pain . ... But of course, the reduction of pain to its physical reality still leaves the subjective experience of pain unreduced' (p. 121). Apparently, there is no way for us to say that the subjective appearance of pain is the underlying reality, given the methodology of reduction.

What exactly are the conclusions to be drawn? From such an argument, we are clearly meant to conclude that a mental property, pain, is distinct from the underlying neurological properties. So, one conclusion we allegedly get is that mental properties are ontologically irreducible to physical properties. However, another conclusion Searle wants from the above is that this irreducibility simply follows from

our methodology of reduction, and further that therefore, there is no threat of property dualism, and no threat to our scientific world view.

Suppose we agree for the moment with Searle, that the above argument shows that mental properties are irreducible. Need we agree that this follows simply from our methodology? I find this expression unclear and slippery. Presumably what Searle means here is that we need not have performed the earlier reductions; we might have agreed with Berkeley, as he says, that the heat was the sensation and not the objective reality (Searle, p. 120). This would show that there is a deep element of convention in our reduction: 'we don't first discover all the fact and then discover a new fact, ...that heat is reducible; rather we simply redefine heat so that the reduction follows from the definition.' One way to interpret Searle here is that given this deep element of convention in our reductions, there is really no significance to reductions at all, since we could easily have redefined the phenomena in another way. This, however, seems a dangerous thing for Searle to be alluding to, given again his commitment to a biological explanation (or, it would seem, to any explanation).

I think his remarks here are a mistake; the point he wants to get across is simply that being a property dualist, and thinking that mental properties are distinct from neurological properties, need not commit one to the view that mental properties are weird, mysterious entities that will forever remain out of the reach of science. However, he himself confuses being a property dualist with being someone who thinks that mental properties are special entities that science cannot explain. Furthermore, he goes so far as to say that the irreducibility of mental properties is a trivial result of methodology. If this is the case, however, we should be able to simply redefine the term 'mental property', in such a way as to deny the import of the Cartesian intuitions. But surely this is not the conclusion that Searle wants. Theorizing is always partly a matter of definition, but this by no means trivializes the results.

I want to suggest at this point that a view of the sort supported by the Lockean intuitions of contingency is much more likely than Searle's to achieve some sort of middle road status between materialism and its opponents. At this point, I want to consider objections to the view that I have been trying to delineate.

4 The new property dualism: objections and replies

The view that has emerged is that certain intuitions of contingency support the view that a mental property such as being in pain is the property of being composed of certain neurological entities that instantiate certain neurological properties. Being in pain is to the stimulation of c-fibers, as being a tree is to the property of having bark and wood in certain organic formations. Particular pains, that is instantiations of the mental property, are made up of instantiations of c-fiber stimulation; similarly, instantiations of the property of being a tree are made of instantiations of the properties of being wood and bark in certain formations. An initial objection involves a worry with the claim that properties are composed of properties. It is clear enough when we say that a thing, like a statue, is composed of matter. But what is it to say that properties are composed of matter and properties? Here I want to distinguish my view from one according to which pain is a complex property with properties as its parts. Rather, on the view I have been suggesting, pain is the property of being composed of certain brain matter, that instantiates certain neurological properties. What I have in mind is simply that the phenomenon of pain, like the phenomenon of lightning, for example, is analogous to the 'phenomenon' of statues and trees, for example. To be a tree is to be composed of a certain sort of matter in a certain sort of formation; to be an instantiation of a mental property is to be composed of a certain sort of matter in a certain sort of neurological formation. This view is put forward as a possibly correct empirical claim of the nature of mental properties, which may or may not be borne out by future theorizing, but which makes some immediate sense of the intuitions of contingency which one may have regarding the identity thesis.

First of all, note that this sort of view can avoid the problems that Searle faced regarding the appropriate construal of the intuitions. This view arises explicitly out of the Lockean intuition, so that what we conceive is that a particular instantiation of a mental property is composed of the underlying matter and neurological properties of the brain. On this view, the causal powers of mental properties are explainable by reference to their composition; but furthermore, mental properties themselves are in a way ontologically reducible to the underlying properties that compose them, contrary to Searle's view. Call this sort of property reduction, POR': a property is said to be ontologically reducible to another if it is found to be (perhaps partly) the property of being composed of matter which instantiates the latter property. Again, mental properties are not strictly identical to the neurological properties; but there could not exist one with the other and vice versa. On such a view, we may then think of mental properties, as Searle seems to want to, as higher-level properties of the brain, properties of being composed of instantiations of lower-level neurological properties of the brain. We would deny that strictly speaking, the phenomenon of pain consists in the stimulation of c-fibers; we would rather say that an instantiation of the property of pain is composed of, or constituted by, instantiations of the property of the stimulation of c-fibers. The relation is not one of identity, since there can be some variation going on in the instantiation of the underlying property while the instantiation of the higher-level property remains constant. We could even go so far as to say that perhaps many cases of reduction that have been construed as cases of identification have really been, in a stricter mode, cases of a similar sort of composition. Take the melting of wax, for example. Rather than claiming that the melting of the wax is strictly identical to the lower level properties involved, perhaps we should say that the melting of the wax is the property of being composed of certain matter that instantiates lower-level properties, i.e. properties of the molecules, their moving faster, the breaking down of their bonds, and so on.

Furthermore, recall the problem Searle had in presenting his view as palatable to physicalists: he went so far as to say that the irreducibility of mental properties is a trivial matter, and as such has no profound consequences for our scientific world view. The problem is that if it is trivial, there is no compelling substantive reason to believe it is true. The view I have been delineating avoids this problem once again. I can cheerfully agree that there is no threat to standard scientific practice posed by my sort of property dualism, even though I distinguish mental properties from the underlying neurological properties. However, I can righteously deny that I am making a trivial claim: my claim is simply that mental properties are strictly speaking not identical to neurological properties. This may not be a terribly exciting claim but it is nevertheless not a trivial one. I can furthermore back up my claim that the view is no threat to science, as I am not one to embrace pure or mixed Cartesian intuitions. Intuitions such as these seem to be more difficult to swallow, from a scientific perspective, than do the Lockean intuitions.

Suppose someone were to object that we might as well simply identify the property of pain with the property of the stimulation of c-fibers. What is there to gain by introducing a clumsy and complicated view that instantiations of mental properties are composed of physical stuff that instantiates physical properties, and that mental properties are the properties of being so composed? What work is done by such a view that could not just as easily be accomplished by the identity thesis? The appeal of materialism as an identity thesis is that it takes some sloppily defined phenomena, our mental life, and discovers a correlation with a neatly defined phenomena, our neurological life, and then claims that the best explanation for the correlation is that what we were thinking of as two distinct properties is in fact one. Perhaps the root of the appeal of a property dualist view is merely some irrational fear of explanatory science, pillaging the last safe haven of mystery.

In reply, I first admit that there is plausibility to the identity view, for just the reasons espoused above. While some might claim that a only property dualist view can account for the essential phenomenality of mental properties, I do not think that it is inconceivable that sensations are neurological properties; so an account of the phenomenal is not the extra work I would claim that my view can handle over the materialist's.

Perhaps, then, there is not a large load of work to be done by this view which cannot be handled by the strict identity thesis. However, there is some work to be done by my view that the identity thesis cannot complete, namely that of making some sense of the intuitions of contingency. The materialist can account for the intuitions by discounting them or explaining them away, or offering some promissory note for the future; but he cannot, I claim, account for them by taking them seriously — i.e. by affirming some construal of them. A plausible construal of their content is that the identity theory is false, and so there really is no way for the materialist to circumvent this problem. Of course the materialist may reply that the intuitions are pre-theoretical and therefore nugatory. I agree that they are pre-theoretical, but disagree that this is a reason to discount them without examining them further. If, upon further examination, we discover an interesting and plausible hypothesis about the status of mental properties, we should be glad that we did not close the door to this discovery earlier. Granted, my construal of the intuitions is such that they are not at all what Kripke had in mind; my view must offer the same promissory note regarding *these* sorts of intuitions. Nevertheless I am able to make sense of the fact that there is some feeling of contingency to the strict identity claim.

To clarify, I do not take it as a constraint of a view of the mind that it must affirm the intuitions in some form. My suggestion is merely that one ought to examine them further to see where they lead, and what implications they may have for further theorizing. If a view affirms some form of the intuitions and still does the important

work of materialism, then this should count as a prima facie benefit of that view. Furthermore, however, it is important to re-emphasize the construal of the intuitions themselves. I am not taking the content of the intuitions to be that I can imagine sensations occurring with any neurological activity at all. I am not taking the content of the intuitions to be that I can imagine mental activity going on regardless of the underlying matter. I am rather construing the intuitions as akin to Lockean intuitions, that I can imagine having the very same sensation even though neuron #5716 fails to fire. I am talking of intuitions of composition, rather than intuitions of radical distinction. The former intuitions, I suggest, should not be as easily dismissed by the materialist. What is suspect about them, that would render them anti-scientific in the eyes of the materialist? They do not seem anti-scientific if we are musing with Locke about trees and their composition; so they should not seem immediately anti-scientific here.

Finally, in response to the materialist's claim concerning the power of theoretical identities, I would say that while theoretical identities are explanatorily powerful, if we can come up with an account of the phenomena that is more complex and yet just as explanatorily powerful, then there is more work to be done to decide which view is more plausibly faithful to the truth. If the truth is that there is a higher level of reality, so to speak, so that there are mental properties that are not strictly identical to neurological properties, then we should strive to work out a view that accommodates this. Of course it is unclear which claim is prior: the claim that reality is multi-layered, or the claim that the identity theory is false and some sort of property dualist theory is true. But the point to keep in mind is that the multi-layered view should not be dismissed immediately.

So, I contend that the property dualist view I have outlined should satisfy the materialist. A more difficult question is whether it would satisfy those who want to take the intuitions of contingency seriously. The answer depends on whether my

construal of the intuitions is a satisfying one to those on the other side. Someone who finds the intuitions plausible may object by claiming that my sort of property dualism still cannot offer a robust account of phenomenality, and as such gives in too much to the materialist. For how could a phenomenal, experiential state, essentially subjective, be identical to anything composed of the instantiation of neurological properties? According to this objection, I have no right to call my view a property dualist view, given the profoundly material nature of mental properties.

I have no qualms with those who prefer not to consider my view a property dualist one. It may be that the view is less of a middle road than a sort of materialism. What makes it a possible middle road is simply that it does take the intuitions of contingency seriously; it is a middle road between those who find something to be said for the intuitions, and those who think that a commitment to the intuitions in some form is tantamount to a denial of materialism. What of the problem of whether I am giving the Cartesian intuitions enough consideration? My reply, as should be clear from my response to the materialist, is that it isn't obvious that a neurological property itself cannot be essentially phenomenal. In other words, I do not think that the construal of the intuitions that such an objector is operating with — the pure or mixed Cartesian, or zombie intuitions — is a plausible one.[24] These are the sort of intuitions, I want to suggest, that could very plausibly fall by the wayside in the face of further theorizing. For we know already that there is some intimate correlation between mental properties and the neurological properties of the brain. This is agreed to on all sides. If those who sympathize with the intuitions will not be satisfied by some view according to which the mental properties are not identical to but are somehow composed of the underlying neurological properties, it seems the only thing that would satisfy them is substance or irreducible property dualism. For we know there is a correlation; if it is not some sort of relation of composition, then the sympathizer is left with mental properties correlated in some very mysterious way with the physical

underpinnings. Perhaps, like Searle, they would want to say that the mental properties are caused by the physical properties. But we have seen that Searle's view can simply rely on the Cartesian intuitions without thereby making itself unpalatable to materialists. The further question to pose to them (and Searle) is what is the nature of those mental properties, if their instantiations are not somehow composed of brain matter and properties? I do not think that such a dualist sort of view is nonsense, but it does seem to go against a naturalistic world view. Rather than embrace such a view whole heartedly, it would be better to go back and review those intuitions we began with: why might it strike one as utterly impossible that, say, a neurological property of the brain be an essentially phenomenal one? Why can't we rather discover that certain neurological properties are essentially phenomenal, or perhaps that they compose essentially phenomenal properties? This could be the sort of scientific discovery of a metaphysical necessity that Kripke himself alludes to in *Naming and Necessity*, akin to the discovery of gold's essential atomic weight. At any rate, I do not expect my sort of view to entice those who construe the intuitions in this radical manner, but I do not find their intuitions to be the sort that one should hold onto tenaciously at this early stage of theorizing. I do expect, however, that my view would satisfy those who find something about the intuitions themselves compelling, but who do not want to be committed to radically dualistic consequences.

If, therefore, one remains steadfast in affirming the plausibility of the Cartesian intuitions, there is nothing I can do to persuade him to my view. But one should be careful to note that in affirming the plausibility and endurance of those intuitions, he is not allowed to claim that mental properties are just like other higher-level properties, as Searle attempts to do. For if, as it may turn out given further theorizing, those Cartesian intuitions are accurate intuitions of some metaphysical distinction, what is being intuited is that mental properties are radically and profoundly different from properties such as liquidity. I admit that it is consistent to suppose that

such Cartesian intuitions be borne out by further theorizing; if this is the case, we will have some reason to suppose that a sort of substance dualism is true. I just do not think that this is very likely.

5 Conclusion

In this paper, I have attempted to delineate a sort of property dualistic view that takes seriously intuitions of contingency of the relation between mental and physical properties, and to show that it is a relatively plausible view. I began by reviewing Kripke's argument against materialism which employs these intuitions, which states that any strict identity statement is necessary and therefore if the intuitions of contingency remain unaccounted for, the materialist identity thesis is falsified. I then went on to explore the possible bases for the intuitions, separating five distinct grounds: pure Cartesian, mixed Cartesian, zombie, multi-realizability, and Lockean. While the first three sorts of intuitions support a claim that mental properties are distinct from the matter than make them up, and in fact do not even supervene on that matter, the Lockean intuitions support the view that while there is some contingency between a mental property and its composition, given that the composition can vary slightly, nevertheless the mental property is determined by its composition.

I then went on to look at Searle's biological naturalism, as a sort of property dualism (Searle's protests to the contrary notwithstanding) that takes seriously intuitions of contingency and attempts to account for them. I argued that Searle's view, as a sort of hybrid view, really cannot be supported by any of the above construals of the intuitions, since Searle wants to hold both that mental properties are caused by physical properties and that mental properties are emergent in the same sense that liquidity is an emergent, upper level property. There is a serious tension between the notion that the Cartesian intuitions show that mental and physical properties are radically distinct, and the claim that mental properties are just like other emergent properties such as liquidity; while liquidity cannot exist if there is no matter

whatsoever, the Cartesian intuitions would have it that the mental can indeed exist regardless of the existence of matter. I further argued that such a view cannot really be as palatable to the materialist as Searle suggests, given his reliance on Cartesian intuitions, and given his emphasis on the claim that mental properties are caused by physical properties. I also examined Searle's discussion of the alleged irreducibility of mental properties, and argued that he was mistaken to try to claim that the irreducibility of mental properties is a trivial result of definition.

Finally, I compared Searle's view to the sort of property dualist view I introduced, that a mental property such as pain is the property of being composed of physical matter that instantiates a certain neurological property, and that instantiations of mental properties are then composed of physical stuff that instantiates that neurological property. I attempted in the final section to show that this view should be palatable to most materialists, as well as to some who find the intuitions compelling. Those who sympathize with the intuitions but do not find the view satisfying are most likely relying on the pure or mixed Cartesian intuitions. The burden is on them to support such radical and pre-theoretical intuitions to a degree which would justify disbelief in some sort of materialism or the sort of property dualism that I have outlined. While I do not take myself to have relied on anything that is wildly controversial, I do think that the results of the discussion point towards something fairly significant: the possible dissolution of the mind-body problem, which makes sense of the intuitions of contingency that linger around the identity thesis.[25]

Endnotes

1. See Kripke's third lecture of *Naming and Necessity*, where he discusses explaining away (certain) intuitions of contingency.

2. For the purpose of simplification, I have in this reconstruction dropped any reference to topic-neutral expressions and properties. The reader should take 'physical expression' and 'physical property' to be abbreviations for, respectively, 'either a physicalistic or topic-neutral expression' and 'either a physicalistic or topic-neutral property'. This should not affect my later points concerning White's argument.

3. In this paper I am using symbols < and > as corner quotes.

4. I think this is the plausible way to interpret A, especially interpreted as a claim about descriptions. If, however, we interpret it so that the meanings of terms can vary across worlds, this would trivialize it, and the effect would be to falsify C. For if we do not hold meanings constant across worlds, then we should draw no conclusions from the possibility of another world, in which they use language differently than we do.

5. A view that Jackson defends, i.e. that the criterion for property identity must be given in terms of predicate synonymy. I would note that discovering the correct synonymies may involve reconceptualization. Further, if one thinks that there is a property for every coherent predicate, one must also admit that there are uninstantiated properties, and so this view will not by itself suffice to propel the view that there must be irreducible (instantiated) mental properties.

6. Presumably this is where someone like Kripke would argue that this cannot be the case, if we cannot explain away our intuitions of contingency. We seem to be able to imagine the sensation remaining the same, while the property of the brain varies, or perhaps if we are more imaginative, disappearing altogether. If I am correct, however, intuitions like these may fall by the wayside in the face of reconceptualization involved in theorizing about the nature of the human mind.

If they do not subside, however, but rather remain stubborn, I would take this to be a serious problem for physicalism. I think the most reasonable route at this stage is to try to follow through on the implications of such intuitions, and see where they might lead, while at the same time we should try to see what it would take to falsify them. What we should not do is feel compelled to deny their plausibility altogether, or to claim that their content follows simply from the meanings of our terms. We need instead to be clearer on what sort of role, if any, such intuitions might play in an empirical theory about the mind.

7. In these papers, the mental properties under discussion are those that are phenomenal.

8. Note that the strength of this conclusion will depend on how we interpret 'at M's disposal'; more on this below.

9. McGinn, however, seems to think it is likely that the mind-body problem is absolutely closed: he 'would not be surprised if it were' (McGinn 1990, p. 16).

10. McGinn seems to conflate the claim that a type of mind would change with the claim that the species would evolve; but it seems that these two should be kept separate, especially considering what he claims about the 'Humean mind', which I will consider later in the chapter.

11. Flanagan also notes that McGinn 'curiously ... fails to explore the prospects of mutual coordination' (Flanagan, p. 18) of the two sorts of data, but he does not take this to be the fundamental difficulty with the argument.

12. Interestingly enough, McGinn seems to propose in a postcript to a later publication of his paper, that he was too hasty in claiming that there are only two ways to discover the property that explains consciousness. He writes 'consciousness too must possess a hidden nature, in which P plays a mediating role' (in Warner and Szubka, eds., p. 115). But in 'The Hidden Structure of Consciousness' (in McGinn 1990) he takes it for granted that his old argument about our cognitive closure with respect to that hidden nature applies. However, it is unclear how his earlier argument is meant to run, given that it relied on a premise which held that there were only two routes to understanding the property that explains consciousness, and given further that we have now been given a possible way out of this dilemma.

13. See Kripke 1972, pp. 144-155. For simplicity, I, like Levine, am assuming familiarity with Kripke's discussion of rigid designators.

14. This argument ploy strikes me as similar to that of the moral relativist, who tries to claim 'murder is terribly wrong, but that is just how it seems to me'.

15. On a related point, Galen Strawson in his recent book is one who seems to take the problem for philosophers of mind to be that of refuting the skeptic. As he puts it, 'if idealism is a genuine metaphysical possibility, then it is arguable that reference to nonmental phenomena may have no necessary role to play in an adequate account of the essential nature of the mental'. If what he means by idealism is simply that it is possible that all of our concepts are wrong, and that hence it is possible that there is no physical world at all, then I would say that nothing follows from this other than the familiar fact that we are fallible. If on the other hand he means that if it is genuinely possible that our minds could exist without physical matter and properties, then this implies the falsity of physicalism, I agree. (Though it may be that what we imagine is the possible existence of something similar to our minds, but the ultimate nature of which is radically different. Perhaps what we imagine is that there could have existed

souls, though these are like nothing that actually exists. This need not have any implications for the truth about our minds, since we are not imagining that our minds could have been such souls.)

16. Brian Loar is another who accepts, as a consequence of his view expressed in 'Phenomenal States', that conceivability does not entail metaphysical possibility. Michael Tye, however, in his recent book, agrees with me (though he uses slightly different terminology) that what seems conceivable is not necessarily genuinely conceivable. However, even he seems to want to claim (pp. 190-193) that some distinction in concepts can explain why there is some apparent possibility that is nevertheless not a real possibility. I would agree with this, though I would note that if we discover that our concepts are implying that some merely apparently possible situation is genuinely possible, that we would find this to be a reason to revise our concepts. That is, on my view, if our concepts implied there was such a distinction, our concepts would then be wrong. He seems to want to admit that our concepts could have such implications, though we would not find this a reason to revise them.

17. The problem is, I take it, the same one that Kripke raised in 'A Puzzle About Belief'.

18. In this paper, I use 'mental property' to mean conscious mental property, i.e. I mean to be discussing those features of our mental lives which are conscious, phenomenal, and experiential.

19. I try to argue for this sort of response in Chapter 2.

20. In 'Troubles with Functionalism', in Block 1980.

21. For simplicity, I am not considering David Lewis's view here that pain is identical whatever plays the appropriate functional role in this world. According to such a view, anyhow, it will be true that pain might have turned out to be something other than it is, where the 'might have' is read metaphysically, and so this would not be a strict identity statement. I take the typical type identity theorist to be asserting a relation of strict identity between pain and c-fiber stimulation.

22. In *The London Review of Books*, May 12 1994, pp. 12-13.

23. Thomas Nagel is one who charges that Searle's biological naturalism is really just a form of property dualism, in his review of Searle's book.

24. This is not to say, of course, that they have never struck me as plausible or as something to be reckoned with. However, on further examination, they do strike me as intuitions that should and will subside once much more evidence is gathered about the nature of the brain.

25. I am grateful to Bob Stalnaker, Alex Byrne, Ned Block, and Ned Hall for helpful comments and criticisms on various drafts of these papers.

References

Block, N. (ed.) 1980. *Readings in Philosophy of Psychology, Volume I.* Cambridge: Harvard University Press.

Churchland, Paul. 1994. 'Betty Crocker's Theory', Review of *The Rediscovery of the Mind,* in *The London Review of Books,* May 12, pp. 12-13.

Descartes, R. 1979. *Meditations on First Philosophy,* translated by D. R. Cress. Indianapolis: Hackett Publishing Inc.

Flanagan, O. 1992. *Consciousness Reconsidered.* Cambridge, MA: The MIT Press.

Jackson, F. 1980a. 'A Note on Physicalism and Heat' in *Australasian Journal of Philosophy,* Vol. 58, No. 1, March.

----------. 1980b. 'On Property Identity' in *Philosophia,* Vol. 58.

Kripke, S. 1972. *Naming and Necessity.* Cambridge, MA: Harvard University Press.

----------. 1979. 'A Puzzle About Belief' in A. Margalit, ed., *Meaning and Use.* Dordrecht: Reidel.

Levine, J. 1983. 'Materialism and Qualia: the Explanatory Gap' pp. 354-361 of *Pacific Philosophical Quarterly, 64.*

---------- 1993. 'On Leaving Out What It's Like' in Martin Davies, ed., *Consciousness: Psychological and Philosophical Essays.* Cambridge: Blackwell.

Loar, B. 1990. 'Phenomenal States'. Page numbers refer to revised copy distributed in Mental Representation Seminar, MIT Philosophy Department, Spring 1994. Original version published in *Philosophical Perspectives,* 4.

Locke, J. 1974. *An Essay Concerning Human Understanding,* in *The Empiricists: Locke Berekeley Hume.* New York: Anchor Doubleday.

McGinn, C. 1990. *The Problem of Consciousness.* Oxford: Blackwell.

----------. 1994. Postscript to 'Can We Solve the Mind-Body Problem?' in Richard Warner and Tadeusz Szubka, eds., *The Mind-Body Problem: A Guide to the Current Debate.* Basil Blackwell.

Nagel, T. 1993. 'The Mind Wins!', Review of *The Rediscovery of the Mind*, in *The New York Review of Books*, March 4, pp. 37-41.

----------. 1979. 'What is it Like to be a Bat?' in *Mortal Questions*. Cambridge University Press.

Searle, J. 1992.*The Rediscovery of the Mind*. Cambridge, MA: The MIT Press.

Strawson, G. 1994. *Mental Reality*. Cambridge, MA: The MIT Press.

Tye, M. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: The MIT Press.

Warner, R., and Szubka, T., Eds. 1994. *The Mind-Body Problem: A Guide to the Current Debate*. Cambridge: Basil Blackwell.

White, S. 1992. 'Curse of the Qualia' in White, *Unity of the Self*. Cambridge, MA: The MIT Press.