Mind-brain identity and evidential insulation

Jakob Hohwy

Monash University

[Published as Hohwy, J. Mind-brain identity and evidential insulation. *Philosophical Studies* 153(3): 377-395. 2011. doi: 10.1007/s11098-010-9524-1.]

1. Introduction

Consciousness is the subject of intense neuroscientific research; it also has an intriguing property of being in some sense private; finally, it is exceedingly hard to see how consciousness could fit into the physical world. I explore how these three facts about consciousness impact on each other.

As the neuroscience of consciousness progresses, it looks very much as if the first major milestone will be identifying the neural correlates of consciousness. I am interested in how this kind of scientific approach can be expected to impact on what we believe about the metaphysics of consciousness. So I want to explore a general epistemic route to a belief about the metaphysics of consciousness whose first stage is finding the neural correlates. I shall endorse the view that this route plausibly leads to belief in the identity of mind and brain.

In particular, I want to factor into this epistemic route what many researchers think is a major obstacle to the scientific study of consciousness, namely the privacy of its subject matter, our phenomenal states. I shall treat this privacy as a non-mysterious kind of epistemic hurdle that I label 'evidential insulation'. This reflects how privacy

is problematic for the science of consciousness and allows us to deal with privacy on a par with other less intimidating epistemic hurdles.

This approach to the subject matter of the science of consciousness changes the complexion of how we should think about the next step, after having identified the neural correlates, and it also changes how we should think about some serious challenges to the idea that consciousness is nothing over and above physical brain states. A core challenge concerns how we can pin down the phenomenal side of the identification without leaving the domain of the physical. This challenge can be captured in an adjusted version of the famous knowledge argument (Jackson 1982, 1986). I show that this version of the knowledge argument turns on making consciousness more evidentially insulated than it really is, which in turn impacts on assignments of degrees of belief in the identity. This proposal differs substantially from the previous and unsuccessful response to the challenge that appeals to different modes of presentation. And it differs from the more recent, hotly debated response to the knowledge argument that appeals to conceptual dualism or phenomenal concepts. In essence, the view is that what drives the knowledge argument is not that there are facts that Mary does not know, but that she lacks a crucial piece of correlative evidence that she can fail to have without the usual objections to the mode of presentation solution and conceptual dualism arising. I thus provide a novel defence of the identity theory.

In Section 2, I set out a generic route to relatively high degree of belief in a posteriori necessary identities. In Section 3 the route is defended against some recent objections concerning reductive explanation. Section 4 sets out the route for the case of identity

of phenomenal and brain states. Section 5 notes how the route is sensitive to matters of evidential insulation. Sections 6-9 assess how the initial high degree of belief in identity can be challenged in ways that suggest it should be adjusted radically down, and this challenge is then answered.

A number of people have taken a roughly similar overall approach. In particular, Jack Smart (1959) early on pressed a scientific take on the identity thesis, which remains central to the debate although it has been interpreted in a number of different ways. Other people have since, also aided by Kripke's notion of a posteriori necessary identities (Kripke 1972/1980), explored this kind of territory. The position defended here differs in various respects, in particular because it does not appeal to different modes of presentation or conceptual dualism, as I shall indicate along the way.

2. An epistemic route to belief in identity: from correlation to identity

To illustrate the epistemic route, consider how the question *"is H_2O=water?"* may be addressed. Assume that we have concepts of H_2O embedded in some theory *T*, and of water embedded in *T** theory. Assume that now the question arises: is water correlated with H_2O (rather than with something else, or nothing at all)? Scientists can use appropriate, theoretically informed methodologies to discern instances of H_2O and of water, and amass good evidence that they occur together. We therefore acquire a high degree of belief that they are correlated.

Now a further question can arise: why is there evidence of H_2O-water correlation? In particular, is it because they are one and the same property, or because they have a common cause, or because one is a cause of the other (perhaps such that water is

irreducible but linked with basic laws to H_2O), or is it divine preordained correlation, or something else altogether?

Assume we get the following kinds of evidence: there is no temporal succession between the occurrences of water and of H_2O, so their relation is not likely to be causal; there is evidence that there is not a common cause since interventions on one (e.g., on the hydrogen bonds of H_2O) invariantly are associated with changes in the other (see Woodward 2003; of course interventions in this sense are designed for causal relations not identities but we can imagine it as a step in a process that rules out common cause and then leads to the discovery that it really is an identity; also see Woodward 2002 for applications of the interventionist scheme to mechanisms); irreducibility would require nomological danglers that integrate poorly with the rest of science (Smart 1959); divine preordination has a very low prior probability and doesn't integrate very well with the evidence of invariance under interventions. These findings tell against the alternative explanations, and are all consistent with identity. Not only are the relevant alternatives to the identity hypothesis poor explainers, it also appears that properties of H_2O can explain some properties of water (e.g., the behaviour of hydrogen bonds under decreasing kinetic energy explains why water expands when freezing; see Kim 2005: Chs 4-5; Block forthcoming, for a discussion of such types of explanation). Given all this, the best explanation of the occurrence of evidence for correlation is therefore the identity hypothesis. This should increase our degree of belief that water=H_2O; that is, that water is nothing over and above H_2O. Though my claim is not that this depicts the actual route to belief in this identity in the history of science, I think it mimics it pretty well. A number of people have suggested roughly this kind of inference to the best explanation to identity, see, e.g., the

discussion in Hill (1991) who has been one of the most forceful contemporary defenders of the identity theory.

3. Challenges to the epistemic route

This epistemic route to high degree of belief that water=H_2O is based on inference to the best explanation of the occurrence of evidence for correlation. Similar accounts (e.g., Hill 1991; Hill and McLaughlin 1999) have been met with the objection that if there really is identity between water and H_2O, then there is no mere correlation. If there is no correlation, then there is nothing to be a genuine best explanation *of* in this kind of case. Therefore we cannot infer to identities as the best explainers of correlation (Block and Stalnaker 1999; Block 2002; Kim 2005, 2008; Block forthcoming). This objection does not touch the above epistemic route to identity. The reason is that the explanandum is not correlation but *evidence* for correlation. It seems entirely in line with common usage to take the notion of correlation in a non-factive sense and to ask what best explains the occurrence of this evidence: identity or a causal relation or something else. The occurrence of evidence for correlation is consistent with there being no real, mere correlation but identity or causation instead.

Block and Stalnaker (1999) themselves stress the use of inference to the best explanation to identity, where the explananda are properties of one side of the identity (water) and the explanantia are properties of the other (H_2O). These reductive explanations are built into the epistemic route I have given above, which is however broader because it also incorporates other pieces of evidence into the inference to identity. That Block and Stalnaker's account can be subsumed under the route is no

surprise, since I treat the evidence for correlation with H_2O as a further to-be-explained fact about water.

The question also arises whether the identity has a genuine explanatory role in the best explanation of evidence of correlation. Kim (2005: Chs 4-5; see also his 2008) objects that an identity serves only as a re-write rule in an explanatory derivation. All the explanatory work is done by the lower level theory (e.g., about H_2O) and the explanation then merely uses the identity to transpose the derived explanandum into the higher level water-terminology. Whether we think such an identity is reductive or not, it doesn't play any explanatory role. If it is not part of the explanation, then it cannot be part of what is inferred to in an inference to the best explanation. It is not a genuine discovery.

Given our epistemic route, it is possible to give the identity a more substantial explanatory role. If we take the identity out, then the bestness of the explanation suffers from more than just an inability to transpose into higher level terminology. This is because the identity is introduced in contrast to other explanations of the evidence: common cause, irreducibility, preordination etc. If we take the identity out, then we lose evidence pertaining to those contrasts (e.g., evidence counting against causation, or common cause, etc). Losing this evidence makes the explanation worse than it would otherwise be. This means that the identity plays an explanatory role and does not just serve as a re-write rule, at least when the inference to the best explanation is set out as above. The difference lies in making inference to the best explanation central, rather than, as Kim, interpreting the explanation merely along the

lines of Nagelian-type deductive–nomological explanation, where bestness considerations and explanatory contrasts play no role.

The proposal gives the identity a substantial explanatory role: it can help best explain something. Conversely, those who believe that these types of identities are a posteriori necessary most often believe that the identity cannot itself be explained (Hill 1991: 24-25; Papineau 1998; Block 2002; Papineau 2002; Kim 2005). This is based on the simple observation that nothing explains why something is identical to itself – it simply is. Adherence to these identities do not deplete our explanatory resources: though nothing explains why H_2O is water, properties of H_2O can help reductively explain properties of water and, via inference to the best explanation, lead to the discovery of the identity (for the use of inference to the best explanation in discovery, see Lipton 2004).

4. The epistemic route to believing that brain states=phenomenal states

We have concepts of brain states, embedded in *P* theory, and of phenomenal states, embedded in *F* theory (e.g., concepts of c-fibres firing in neurophysiology or perhaps of a particular type of thalamocortical oscillation, and of pain or colour perception in folk theory). First, the question arises: are phenomenal states correlated with brain states (rather than with other states, or with nothing at all)? We use appropriate, theoretically informed methodologies to discern instances of brain states and of phenomenal states, and amass evidence that they occur together. We therefore acquire a high degree of belief that they are correlated. This is what researchers endeavour to do in the search for the neural correlates of consciousness (cf. Chalmers 2000).

Now a further question arises: why is there evidence of correlation between brain states and conscious states? Again, this question is approached through explanatory contrasts between various competing hypotheses: identity, common cause, irreducible causal relations, preordination. As in the case of H_2O=water, there is evidence against the alternatives and evidence in favour of the identity hypothesis that therefore transpires as the best explanation of the correlation evidence. Our degree of belief that brain states=phenomenal states should therefore increase (this should of course be spelled out for contrasts among distinct types of phenomenal states). As above, the identity itself is an explainer of the occurrence of correlation evidence, and is itself left unexplained.

What makes the identity the best explainer? Against causality (in the form of basic psychophysical laws), the cotemporality of the correlated states, and the poor ability for integration with the rest of science of nomological danglers relating large assemblies of neurons with simple phenomenal states (cf. Smart 1959). Against common cause explanations, the invariant relation between the correlated states under intervention. If epiphenomenalism is seen as a contender, we can cite the evidence for mental causation and in general the difference and similarity relations among phenomenal states (Shoemaker 2007); if interactionist dualism is seen as a contender, we can cite the evidence for the causal closure of the physical, and also Kim's (1998; 2005) causal pairing problems with accounting for mental causation for dualism. Against preordination, very low prior probability, and the invariance under intervention too, since it is not predicted by the hypothesis of divine intervention. In addition, properties of brain states can reductively and contrastively explain relational

and structural properties of conscious states, just as properties of H_2O could explain properties of water (for discussion, see Hohwy and Frith 2004).

This seems to me a fairly convincing epistemic route to assigning a high degree of belief in the a posteriori necessary identity of brain states and phenomenal states (for discussion and different variations on the route, see Smart 1959; Hill 1991; van Gulick 1993; Melnyk 2003; Polger 2004). My twist on this familiar story, as above, is that I give a central role to inference to the best explanation of the occurrence of (non-factive) correlation evidence, and I am able to give a substantial explanatory role to the identity, rather than having it as a mere re-write rule.

At this stage I am fairly perfunctory in inferring to identity rather than dualism or the other non-identity candidate explainers. I do not want to suggest that this is the end of the mind-body problem; in particular, I have ignored what many think detract decisively from the bestness of the identity thesis as an explanation, namely that it seems necessarily false. This is what the famous anti-physicalist arguments, including those discussed by Descartes, Kripke, Jackson and Chalmers (and others), are often taken to show. I shall discuss how the route to identity I have set out may be challenged by such arguments in Section 6. But leaving aside those arguments for the moment, the inference to mind-brain identity seems not that different from the types of inferences one would make for cases such as water and H_2O, lightning and certain types of electrical discharge, and so on. In those cases explanations involving identities fare better on the gamut of bestmakers than explanations involving other relations between the correlated properties, and for pretty much the kinds of reasons I have gestured at above.

Before I enter the familiar anti-physicalist debates about these identities, I consider some properties of the epistemic route to belief in identity.

5. How the epistemic route to identity is sensitive to evidential insulation

This will be a fabricated example about some elementary particles. It will somewhat stretch epistemic credulity but this will not be relevant for the moral.

Assume we have concepts of *snark* particles, embedded in what we will call *LC* theory (for *L*ewis *C*arroll) and concepts of *boojum* particles, embedded in *LC**. Now the question arises whether snarks are correlated with boojums rather than some other particles. We use the available theoretically guided methodologies to discern occurrences of snarks and boojums and discover that they coincide. Consequently, and as in the case of evidence for H_2O-water correlation, our degree of belief in their correlation goes up.

Next, the question arises: Why is there this occurrence of correlation evidence? As before, this is approached through contrastive questions: is it identity rather than common cause, or rather than preordination, etc? The identity of snarks and boojums transpires as the best explanation of the evidence: snarks have certain properties that can explain some boojum properties and there is evidence against common cause, preordination and the rest. Consequently, and as above, our degree of belief in snarks=boojums goes up.

Now we complicate the story: assume that boojums are relatively *evidentially insulated*. That is, we acquire evidence for their occurrence *only* by observing a certain kind of vapour trail (this is where epistemic credulity is stretched since normally we think a theory like *LC** would be associated with some additional sources of evidence for their existence; but I ignore this here). Evidential insulation concerns the number of methods for acquiring evidence for the occurrence of something. The more such methods, the less the evidential insulation. If there are no methods the insulation is total. (Snarks, in contrast to boojums, are not significantly evidentially insulated, imagine we have a decent range of standard scientific methods for discerning their occurrence).

Someone now raises the possibility of *vacant trails*, trails without a particle (the possibility could also concern some kind of gerrymandering of pairings of trails and particles, this doesn't raise immediate additional issues so I set it aside until I need it, in Section 9). Assume that there is no new empirical evidence that suggests this possibility – it is based merely on the fact that nothing a priori excludes this possibility. Then, given that we ascertain the occurrence of boojums only by observing vapour trails, this possibility is a serious threat to our methodology. The possibility introduces a context in which, to be confident in how we assigned degree of belief in the identity, we need additional information about the frequency of vacant trails. In particular we need to eliminate the possibility that vacant trails are very frequent: if they are, then the correlation evidence we were trying to explain is spurious, and identity is not the best explanation of spurious correlation evidence. The problem is that we cannot rationally use the vapour trail method to validate itself, and since boojums are evidentially insulated there is no other method available. That is,

our available information does not after all allow us to rationally decide between snarks being identical to boojums or not. Therefore our degree of belief in the identity should decrease again.

The result is a looming epistemic gap between snarks and boojums: we are not after all justified in holding that boojums are nothing over and above snarks. However, this looming gap from vacant trails depends in the first instance on the evidential insulation of boojums. It wouldn't loom if there were further methods for ascertaining the occurrence of boojums because those methods might be untouched by any particular a priori doubt. And, of course, snarks might *be* boojums all the same.

This epistemic notion of evidential insulation, and the gaps it can give rise to, is exploited below.

6. The challenge: explaining the contingency while remaining physicalist.

One of the most influential challenges to physicalism about consciousness is the knowledge argument (Jackson 1982; Jackson 1986). This argument is directed against a priori physicalism, according to which truths about phenomenal states are entailed by truths about physical states together with a priori conceptual analysis. The core of the argument is that the super-intelligent colour scientist Mary, who is trapped in a black and white room, will fail to appreciate all the truths about colour perception as long as she knows only all the physical truths and haven't experienced colour herself.

The argument is not directed against a posteriori physicalism of the sort outlined above. Therefore, some believe that opting for a posteriori physicalism is a way to

avoid the knowledge argument (there are many in this camp, see, e.g., Loar 1997; Block 2002). Defenders of a posteriori physicalism use the aposteriori part of their thesis to suggest that the apparent contingency of the mind-brain relation is indeed only apparent. This must however be done such that it is clear that the mind-brain identity is retained (see, e.g., Hill 1991; Papineau 2002). There is much debate about the successfulness of such attempts at appealing to the necessary aposteriori to account for the appearance of contingency while trying to remain physicalist (for useful overviews, and arguments, see Chalmers 2002; Chalmers 2006). Here I devise a new such attempt. For this, I update an important early objection to the identity theory and present it as a kind of knowledge argument that directly engages the kind of epistemic route to belief in a posteriori necessary identity that I have pleaded for above. In my view, a very natural interpretation is that the threat to the identity theory from this version of the knowledge argument is innocuous since it turns merely on making the evidential insulation of phenomenality worse than it really is. In this way, the inference to identity as the best explanation of the correlation suffers because a crucial piece of correlative evidence is lacking. However, as I shall show, this evidence can be lacking without the usual objections arising

An early and important objection to the identity theory was the one Smart attributed to Max Black concerning how the phenomenal side of the identification is "pinned down" (1959: 148-50; 'pinning down' is Smart's placeholder for an appropriate semantic notion; Block 2006 has recently directed renewed attention to this objection). This is an intricate objection. Phrased in contemporary terms it aims to show that a posteriori physicalism cannot remain a posteriori *and* physicalist on any straightforward conception of how the phenomenal states are pinned down. If the

means by which they are pinned down are non-physical in the sense of being something over and above the physical, then physicalism is obviously false. If the means are topic neutral (as Smart argued), then it seems to me we turn the theory into an a priori physicalism, since topic neutral descriptions look very much like the ramseyfied functionalisations characteristic of analytical functionalism. Lastly, if the means by which the mental is pinned down are physical, then we also lose the a posteriori element because then it is a foregone (i.e., a priori) conclusion, that the mental is the physical. For example, let us say we pin down "neural" property in virtue of its being thalamocortical oscillation and pin down "phenomenal" property in virtue of its being thalamocortical oscillation. Then there is no substantial, empirical question about whether "neural" and "phenomenal" properties are identical. The way we have pinned them down ensures that they are. Moreover, this seems inadequate to ensure phenomenal states have been pinned down at all, so it would be reasonable to doubt that the mind-body problem had been solved this easily.

The challenge is to show how the means of pinning down genuine phenomenal states can be consistent with physicalism without losing the a posteriori element. This is an extremely difficult task for the a posteriori physicalists because it seems to require them to tell a paradoxical story on which the physical is non-physical (which is why Smart opted for topic-neutrality). Notice that the structure of the challenge is conveniently close to the Mary-scenario: even with all the physical knowledge it seems it is not possible to pin down the phenomenal states for the purposes of an inference to a posteriori mind-brain identity, and if non-physical knowledge is required to pin them down, then physicalism is false. It will give some room for

manoeuvre if the challenge is set out in these more epistemic terms. I thus begin by developing an appropriately adjusted version of the knowledge argument.

To distinguish Jackson's original target and our case of a posteriori physicalism, we call the hero of our story Baker. Baker's task is to go through the epistemic route to identity that I have set out. He must then begin by amassing evidence for mind-brain correlation. To begin doing this he has to be able to pin down occurrences of brain states and phenomenal states. It is stipulated that he has all the physical information inside the room so the brain side of things should not be a problem. What about the phenomenal side of the identity? As with Mary, Baker is prevented from having some kinds of phenomenal experience, for example of pain or colour. He must then rely on introspective reports of such states from other people since such reports are behavioural and thus admissible inside the room. Once the evidence is in he can consider, purely on grounds of physical evidence, what is the best explanation of the evidence. As set out above, it seems he can rationally end up by assigning a high degree of belief in the mind-brain identity.

For our version of the knowledge argument to work, something about the possibility of gaining new knowledge when released from the room must prevent Baker from rationally arriving at this conclusion while still in the room, just as something prevents Mary from rationally deducing all the phenomenal facts (e.g., the notion that qualia concepts cannot be functionalised and therefore cannot be captured before she leaves the black and white room). As stipulated, it cannot be something about the evidence for the occurrence of brain states so in Baker's case it must concern the use of introspective reports to pin down the phenomenal states. However, it cannot

concern the behavioural aspects of amassing these reports, for that is just more physical information. Instead what prevents the inference to identity must concern the very method of using introspective reports to pin down the phenomenal states. In order for Baker's own phenomenal experience (which he gains on being released from the room) to be relevant here, this version of the knowledge argument must seek to undermine the use of introspective reports. That can be done by saying that, from inside the room, Baker cannot rule out that the method is unreliable, that is, that it fails properly to pin down phenomenal properties. This version of the knowledge argument works by pointing out some doubt about the reliability of introspective reports that can arise when Baker's own phenomenal experience is left out of the story. It works, in other words, by pointing to how Baker from inside the room cannot rationally rule out the possibility of *vacant reports*: spurious introspective reports that do not originate in phenomenal states.

Baker's predicament is that when the possibility of vacant reports is raised he is put in a context of doubt about the reliability of the introspective reports he has access to from within the room. He cannot validate those reports from within the room by asking people outside the room about their experiences because the method of sampling introspective reports cannot be used to validate itself. But now it seems a gap has opened up between brain states and phenomenal states. Baker has no resources, starting from the physical side, for rationally assigning a high degree of belief to the conclusion in favour of mind-brain identity.

It seems the only way to gain independent hold of the method of using introspective reports is to support it indirectly via introspection of one's own phenomenal states,

and then arriving at justified beliefs about other's phenomenal states by applying the argument from analogy (inferring to other's phenomenal states via relevant similarities in respect of behaviour and physical realisation). However, this is precisely the kind of additional evidence that is not available inside the room, so Baker can't close the gap. Once Baker leaves the room he will be able to close the gap and re-instate his confidence in assigning a high degree of belief in the identity (assuming that he discovers that introspective reports are really reliable enough).

Against our kind of a posteriori physicalism, this version of the knowledge argument therefore proceeds by imagining phenomenality to be even more evidentially insulated than it actually is. It is no surprise that such a gap arises when some property is stipulated to be evidentially insulated. But it is entirely plausible to say that it is an artificially forced gap, which, just as in the snarks=boojums case, is easy to bridge if only further methods are available – such as introspection. In the snarks=boojums case it would be too hasty to draw the conclusion that boojums are something over and above snarks, and, likewise, in the brain states=phenomenal states case it would be too hasty to conclude that phenomenality is something over and above brain states.

How does this tie up with the available moves vis-à-vis the original knowledge argument? This take on the version of the knowledge argument can be phrased in terms of epistemic contextualism: Baker has all the pieces of physical information (all the physical truths) inside the room but is not entitled to assign a high degree of belief in all of them, and hence he cannot be said to know them. As an argument against a posteriori physicalism, the argument is therefore self-undermining: the physical knowledge that he initially has *eludes* him (in Lewis' (1996) term) when the doubt

about the correlation evidence and the introspective method is introduced into the context via the notion of vacant reports. This keeps faith with the way the knowledge argument is normally presented; in particular, we can accept the first premise that Baker has all the physical knowledge. We just add that the conclusion doesn't follow because the context changed midway through the argument when a new possibility is introduced that Baker cannot rationally eliminate from inside the room.

Others have argued that Mary may not have all the physical knowledge but mostly this is argued by pointing to some class of properties (e.g., intrinsic properties) that her theoretical physical framework simply cannot capture (Stoljar 2006 is the strongest defender of such an ignorance view; for discussion see Hohwy 2005). Here I treat knowledge as something like justified true belief and note that, for our case of Baker, some of his true beliefs about the physical lose their justification once the possibility of vacant reports is introduced into the context. The ignorance induced in this way is less damaging than the standard appeal to ignorance because the knowledge was had at one point and then lost, rather than never had in the first place.

With this we can move to the final consideration in this section, concerning the explanation of the appearance of contingency. Water is identical to H_2O, and yet it seems as if we could have discovered that water is not H_2O. Following Kripke (1972/1980), many explain that what we could have genuinely discovered is not that *water* is something different from H_2O but that what *seems* like water (the *watery stuff)* is. It is only when we confuse the notions of *water* and *watery stuff* that the identity seems contingent. While in the grip of the confusion, one thinks that the identity is contingent because an epistemic possibility has been pointed to in which water fails to

be H_2O. Many also follow Kripke in holding that this explanation doesn't work in the mind-brain case because what *seems* like a phenomenal state *is* that state – what seems like pain is pain. Since my approach concerns the justification of what is believed and not the content of belief, there is no such proposition confusion here. My explanation of the appearance of contingency works by pointing to an epistemic possibility, namely in which there are vacant reports. As long as phenomenality is evidentially insulated this possibility cannot be ruled out, and it should not be concluded that brain states are identical to phenomenal states. This possibility is consistent with brain states in fact being identical to phenomenal states, so the contingency can be merely apparent. The epistemic possibility rests on presenting phenomenality as more evidentially insulated than it actually is. Water, boojums and phenomenality seem on a par in this respect, in particular there is nothing about phenomenality that suggests it cannot be presented as more evidentially insulated than it is. So this general type of strategy does seem to work for the mind-brain case too, in contrast (as many would have it) to the proposition confusion strategy.

We can then summarise Baker's "epistemic journey" like this. Best explanation can lead to discovery (e.g., the red shift in the light spectrum of some stars can be explained in terms of recession, and can thus lead to discovery of recession; see Lipton 2004). In the case of Baker, best explanation leads to the discovery of the identity of brain states and phenomenal states. However, Baker doesn't properly discover or learn this until he comes out of the room and can take the method of own introspection in possession. For it is only when he can introspect that he can rationally iron out some doubts that prevented him from making this discovery inside the room.

7. Does the gap re-open?

Here is an objection to the view I am proposing. No matter how evidentially insulated boojums are, it is always possible to raise the possibility of vacant trails and, similarly, no matter how evidentially insulated phenomenality is, it is always possible to raise the possibility of vacant reports. That is, even if there were a multitude of methods to gain evidence for the occurrence of boojums or phenomenality, epistemic gaps induced by the possibility of vacant trails or vacant reports or whatever could always open up. Then it is not evidential insulation per se that produces the gaps. But if the degree of evidential insulation makes no difference to the occurrence of the epistemic gap, then the proposed defense of a posteriori physicalism in terms of degree of evidential insulation fails. In particular, it is proposed that the gap is closed once the evidential insulation of phenomenality is lessened but why can't it just open up again in spite of this new degree of insulation? However, this objection can be met. My proposal is consistent with the idea that epistemic gaps can occur for any degree of insulation, n (with n=0 being highest degree of insulation). Of course, we would say that the probability of a gap decreases as n goes up: the more ways one gets evidence for the occurrence of something the less likely it is that there is error, and if n=0, then it is very likely that we are wrong (the relationship would not be linear as going from n=0 to 1 is extremely important, the next couple of degrees being quite important too, and then, after n=5 perhaps, the importance of decreasing insulation would presumably taper off). The thrust of my proposal is that the antiphysicalist argument works by making phenomenality appear more evidentially insulated than it really is – it takes n from 2 down to 1. Thereby it creates the "doxastic illusion" that a gap is very probable, when in reality (with n=2) it isn't very probable.

Having said that, it is true that in principle a new gap could open up, even with *n*=2, or indeed for any higher *n*. But notice that it is not true that epistemic insulation is irrelevant to such gaps. On the contextualist view that I am presenting, the gap does genuinely close when a new method deals with doubts about the old method. As the context changes new doubts can arise that target the latest method and this doubt can only be conquered once this new level of evidential insulation is broken. It is the evidential insulation at any level that makes the gap possible, so the insulation is relevant after all. In this respect mind-brain gaps are no different from the kinds of gaps that can in principle open for water and H_2O or that did open in my story about the boojums. Thus what I am arguing here is that the mind-body gap is an innocent type of epistemic gap that can occur all over science, and which can be closed and which can re-open just as innocently. If I am right, then we can appeal to introspection and thereby close the gap just as efficiently as the gap can be closed in the case of boojums when a second method is found, and just as efficiently as it is in fact closed in the case of water and H_2O. Our belief in the mind-body identity is firm but not unshakeable, just as our belief in water=H_2O is firm but not unshakeable. What our version of the knowledge argument does is that it makes the mind-body identity appear very shaky because it makes the insulation more severe than it really is.

In the water-case, a gap tied to water's current level of evidential insulation could in principle open up. But we would find it hard to take such a challenge seriously since the evidential insulation is very little as it is. That is why I had to tell the story in terms of boojums where our intuitions about the science are not so ingrained. The question arises what would make a gap open up in the mind-body case, even after we have re-appropriated the method of introspection. The first thing to notice is that our

version of the knowledge argument cannot make *that* gap open up precisely because this argument prevents us from appealing to the method of introspection. A different argument is needed according to which even the deliverances of the method of introspection could be vacant. A quick response to this scenario is then to say that I am here only concerned to defend a posteriori physicalism against the appropriate version of the knowledge argument, setting aside the issue of whether other arguments could pose further threats.

Nevertheless, we can speculate about what would lead us to suspect a re-opening of the gap. One way to go would be to raise the possibility that an introspective experience as of pain can fail to really be an experience of pain. This will have to be controversial given the Kripkean point that there is no is-seems distinction for phenomenal states. There is however lively debate about whether introspection as such is reliable (Schwitzgebel 2008). If the case can be made, then the gap that opens will be difficult to close since we have precious little idea about what an additional, third method for accessing phenomenality would be like.

Another way of re-opening the gap could be to generate the doubts about introspection indirectly, via an attack on the argument from analogy as a solution to the other minds problem. Baker needs to get out and have experiences himself in order to close the epistemic gap and rationally form the belief in mind-brain identity. Crucially, he engages the argument from analogy in order to judge about the reliability of other's introspective reports. The argument from analogy is little help if the creatures in question are fundamentally different from Baker in their physical realisation (Hill 1991: Ch 9; Block 2002). So whereas the gap between mind and body

can be closed for creatures like us, Baker cannot rationally close it for other creatures. We are left therefore with something like Block's (2002) "harder problem of consciousness": that we have no conception of how to address the question of whether such physically different creatures are conscious or not. On my account, the nature of this problem is clear. The appropriate version of the knowledge argument relies on increasing the evidential insulation of phenomenality, with rational belief in mind-brain identity being restored once insulation is decreased by re-appropriating introspection. In contrast, the harder problem of consciousness relies on the fact that the method of introspection cannot be used to decrease evidential insulation of the potential phenomenality of fundamentally different creatures. And, we currently have no idea what else we could use to decrease the insulation.

All this goes to show that even if the means by which Baker pins down the phenomenal are indeed physical, closing the gap is by no means a trivial matter: new doubts can arise that can undermine the belief in identity. Now, importantly, it must be shown that the introspective evidence that enables Baker to close the gap (in his current context) is not something over and above the physical. I do that by comparing to other, more problematic approaches.

8. How the appeal to evidential insulation differs from appeal to different modes of presentation.

A common response to the knowledge argument appeals to different modes of presentation (see, for example, Churchland 1985). The basic idea is to notice how it doesn't follow from the fact that something new is learned that it is knowledge of a *new* fact. The reason is that it could be new knowledge of an *old* fact, which is now

presented under a different mode of presentation than before. On such an account Baker cannot even formulate his research project because he cannot pin down the phenomenal side of the identity until he gets out and appropriates the new mode of presentation.

The text-book problem for the strategy that appeals to different modes of presentation is that it requires that what is referred to possesses a new property that allows it to be picked out as the referent under this new mode of presentation. But if this property was not known in the room then it is a non-physical property. And then physicalism still cannot be true (Braddon-Mitchell and Jackson 2007: 137).

On my account, it *is* possible for Baker to formulate his research project while in the room for he may possess the requisite modes of presentations. The problem for him is rather that he cannot rationally assign degrees of belief in identity, for the belief is conditional on some evidence that he is not rationally entitled to rely on.

However, an objection could be that my account is very similar to the appeal to different modes of presentation since getting access to an additional method for sampling evidence about something is similar to getting access to an additional mode of presentation. And then the same problem could arise that the property that is revealed by the new method is a property that should have been known inside the room, if really physical.

My initial response is this. On the modes of presentation account, Baker has only the neuroscientific mode of presentation of phenomenal states while in the room whereas

on my approach he has in addition the introspective reports of other subjects. This is why initially he can formulate his research project and then assign a high degree of belief to the identity. Plausibly, the kinds of property that are sampled by introspective reports are the very same kinds of property as are sampled in one's own introspection – namely phenomenal properties. So no new property is introduced when Baker leaves the room, just a new method for sampling the same property (on the assumption that the vacancy doubt about the reports is unfounded). This is analogous to how use of an optical microscope and an electron microscope could sample the very same property even though they are different methods. If doubts arose about the evidence obtained by using the optical microscope, then we would have to await access to the electron microscope in order to firm up our assignment of degrees of belief.

This suggests that Baker's doxastic situation on my account and on the modes of presentation account are quite different. However, there is a deeper worry. If all Baker needs is information about the reliability of the method of using introspective reports, then that information can hardly be physical if he doesn't have it while in the room. If it is not physical, then physicalism is false for then he should not assign a high degree of belief to the identity on the basis of having only all the physical information.

To answer this objection, I wish to take the a posteriori, scientific element of the story very seriously. Let us say that one of Baker's black and white neuroscience textbooks states "introspective reports of colour experiences are reliable" and perhaps that it spells this out in terms of the neurophysiological evidence concerning the properties that constitutes the ability to introspect and how reliably they connect with the neural

correlates of phenomenal states of colour experience. Couldn't this assuage the doubt and make him be confident in assigning a high degree of belief in the identity? If the answer is 'yes' to this question, then I just belie the intuition that he is genuinely barred from assigning a high degree of belief inside the room. There is though a way to answer 'no' to that question. The doxastic steps he would go through are like this. He is in a context of doubt about the reliability of one of his methods, that of sampling phenomenal states through use of introspective reports. When presented with the true textbook statement that the method is in fact reliable he should, given his current context of doubt, rationally ask what the evidence is for this very statement. There seems to be only two possibilities: either it is evidence that relies (at least in part) on other subjects' introspective reports, or it is evidence that relies (at least in part) on his own introspection. There is no third option, due to the evidential insulation of phenomenality. It cannot be the latter because he hasn't had the requisite introspection yet. And he cannot rationally rely on the former for that is based on the very method that he is in the process of evaluating. The statement that the reports are in fact reliable may very well convey some physical information (and must, if physicalism is true) but, due to the way we rationally assign degrees of belief, that piece of evidence is inadmissible. He must wait until he himself gains access to the method of introspection before he can re-assign a high degree of belief in mind-brain identity.

9. How the appeal to evidential insulation differs from appeal to conceptual dualism.

A more recent response to the knowledge argument appeals to conceptual dualism, which also deals with the knowledge argument by noting how it doesn't follow from the fact that something new is learned that it is knowledge of a *new* fact. The reason is

that it could be new knowledge of an *old* fact, which is now pinned down using a newly acquired set of *phenomenal concepts*. Again, on such an account, Baker cannot even formulate his research project because he cannot pin down the phenomenal side of the identity at all until he comes out and acquires the phenomenal concepts.

This strategy appeals to the particular nature of phenomenal concepts (concepts that Mary, and Baker, lack while in the room), and it is held, for example, that they are recognitional or perhaps quotational (e.g., Tye 1995; Loar 1997; Hill and McLaughlin 1999; Block 2006). The modes of presentation strategy appeals to a general idea of different modes of presentation, whereas in contrast, the conceptual dualism strategy appeals to the notion that there is something special about phenomenal concepts. This specificity should inoculate it against the problems that beset the modes of presentation strategy. In particular, if phenomenal concepts pick out phenomenal properties somehow directly, without the mediation of a mode of presentation, then this approach can avoid the problem had by the modes of presentation approach. The hope is then that, if phenomenal concepts are themselves necessitated by the physical (Chalmers 2006; Levine 2006 both have their (in my view legitimate) doubts about this), then this could explain the appearance of an epistemic gap while allowing that there never was an ontological gap.

The question arises whether the evidential insularity approach needs to appeal to something like phenomenal concepts given that both look like alternatives to the modes of presentation approach. My answer to this will be that the evidential insulation approach only needs to appeal to the possible availability of further, independent evidence and that, on some accounts, one but not the only possible route

for acquiring such further evidence goes through possession of phenomenal concepts. In so far as the story about phenomenal concepts helps explain Baker's doxastic situation it is in virtue of how it affects the quite general story about belief formation, not in virtue of anything special about phenomenal concepts.

If the conceptual dualism approach is on the right track, then it is the lack of phenomenal concepts that explains Mary's epistemic predicament. But as Stoljar (2005) shows, the knowledge argument can be given in a version with Experienced Mary who was let out of the room and thus acquired phenomenal concepts, and who was then put back in the room again but with amnesia concerning what kinds of situations give rise to what phenomenal experiences. Stoljar argues that, even though she has the phenomenal concepts, there are truths about the phenomenal that Experienced Mary cannot deduce such as whether seeing a Red Delicious apple gives rise to red experiences or nausea experiences. This shows that phenomenal concepts are irrelevant for explaining the appearance of an epistemic gap.

Stoljar's finding is useful to mark a substantial difference between the evidential insulation approach and the conceptual dualism approach. Thus consider Experienced Baker who is back in the room after acquiring phenomenal concepts. He has amnesia for all the evidence about the occasions on which the phenomenal concepts are correctly applied. But this is exactly the kind of evidence needed to overcome the doubt about the reliability of the introspective method. Experienced Baker is in a context of doubt where he needs to garner independent evidence that when people report, for example, that they have a red experience on seeing a Red Delicious apple they are by and large right.

There might however be a way to give phenomenal concepts a role for garnering the right kind of evidence. It may be that Experienced Baker can be truly said to still possess the phenomenal concepts when back in the room only if he hasn't amnesia for *all* his past phenomenal experience. We should then perhaps allow that he is still able to remember what it was like to have experiences of red. So, according to the conceptual dualist, when Experienced Baker employs the concept in, for example, remembering what it was like to experience red, then, in doing so, he somehow directly tokens a red experience. In such a case he could perhaps, via the argument from analogy, use that experience as the required independent evidence that introspective reports of colour are by and large not vacant. This is progress but there is a limit to the doubts he can counter in this way. He cannot use this method to deal with the different but just as serious doubt that introspective reports are non-vacant but somehow gerrymandered (e.g., reports of experiencing red are paired with all sorts of phenomenal states): to deal with this doubt he needs to overcome the amnesia for the episodes in which phenomenal concepts are correctly applied. The gerrymander doubt is serious too because it also undermines the evidence delivered by studies of the neural correlates of phenomenal experience. So the difference between the evidential insulation approach and conceptual dualism approach remains: even with the concepts, Experienced Baker still cannot rationally believe in the identity.

Notice, also, that here possession of the phenomenal concepts helps against Baker's epistemic predicament concerning vacant reports only because they bring in their wake some of the evidence he needs. This suggests that what best explains why

possession of phenomenal concepts could possibly begin to account for the appearance of contingency of the identity is that it begins to decrease evidential insulation. That is enough for my account to get off the ground and it is a story that can be told without invoking any specific story about phenomenal concepts: we could just allow Experienced Baker to remember what it was like to see red.

Consider now Chalmers' (2006) criticism of conceptual dualism (see also Levine 2006). This presents a dilemma (not unlike Max Black's original objection that I canvassed above) for the a posteriori physicalist who believes that appeal to phenomenal concepts can explain the appearance of contingency of mind-brain identities. Either the phenomenal concepts themselves can be accounted for as physically constituted, or they can't. If they can, then a new gap opens up for how can an entirely physical story about phenomenal concepts entail that Baker tokens *phenomenal* states directly when applying those concepts? The problem here is that just as the physical story is, in Chalmers' phrase, too "tame" to account for phenomenality in the first instance so it is too tame to account for why phenomenal concepts are *phenomenal* and thus to account for the appearance of contingency. On the other horn of the dilemma, if phenomenal concepts cannot be accounted for in physical terms, then physicalism is just false.

I think this is an impressive challenge for conceptual dualists. If I can show that my account is not threatened by a version of this challenge, then I have not only defended it against a serious potential challenge, I have shown that it likely differs significantly from conceptual dualism. Given that the account is intended to be physicalist, I need to dodge the first horn of the dilemma. The question is therefore how the existence of

the introspective method of sampling phenomenal states can be accounted for in physical terms and still explain the appearance of contingency that gives rise to the epistemic predicament. In terms of Baker: shouldn't he already know, within the room, that the introspective method is reliable, if it is?

My response here is the same as in the case of the modes of presentation approach and again relies on taking the a posteriori, epistemic aspect of the story very seriously. Baker is indeed presented with the (as it happens) veridical physical information that the introspective method is reliable. To the extent this story concerns phenomenal states it invites him to identify those with physical states, as per our epistemic route to identity. But, he finds himself in a context of doubt about the information, in particular he comes to suspect that other people's introspective reports are vacant or gerrymandered. As we, from an a posteriori physicalist perspective, have set up the task for Baker, the neuroscientific story about the reliability of the report must itself rely in some part on evidence for the correlation of brain states and phenomenal states. This evidence therefore employs the very method currently being doubted and he cannot rationally rely on this evidence even though we may assume physicalism is true and that it is in fact veridical. Since he cannot use the method to validate itself the only way to rationally overcome this doubt is to decrease the evidential insulation by himself having enough phenomenal experiences to enable him to test, via the argument from analogy, whether or not introspective reports are vacant or gerrymandered. Chalmers' point, as applied to Baker, was that if the story about the reliability of introspective reports is indeed a physical story, then it is too tame to explain why it is introspective reports about *phenomenal* states, and thus it cannot explain the appearance of contingency. What I have just shown is that if the story

about the reliability of introspective reports is indeed a physical story, then that story can explain why they are reports about phenomenal states: this is itself part of the evidence that goes into the explananda in the inference to the best explanation to identity. Therefore phenomenal states are retained in the story and, crucially, the epistemic gap remains even after Baker is given this true physical story, for reasons having to do with rational belief formation.

10. Concluding remarks

Finally, I return to the three aspects of consciousness that I mentioned at the very start: there are intense scientific efforts to reveal its neural correlates, it is private, and it is hard to fit into the physical world. In my discussion I have played these three aspects against each other and my final account is constrained by them. I gave a general and rather unmysterious epistemic account of the privacy in terms of evidential insulation and used this in a novel way to leverage rational belief in the identity of phenomenal states and brain states, based in the neuroscientific evidence. To me, this is pleasing since I am primarily interested in the mind-body problem as it in some sense relates to epistemic and scientific concerns – to what we should believe about phenomenality.

Acknowledgements

Thanks for comments on an earlier draft to the audience at a seminar at Monash University, to Jesper Kallestrup, and to reviewers. This research is partly funded by Australian Research Council Discovery grants DP0988514 and DP0984572.

References

Block, N. (2002). The Harder Problem of Consciousness. *The Journal of Philosophy* XCIX(8): 1-35.

Block, N. (2006). Max Black's objection to mind-body identity. *Oxford Studies in Metaphysics*. D. Zimmerman. Oxford, Oxford University Press. II: 3-78.

Block, N. (forthcoming). Functional Reduction. *Supervenience in Mind: A Festschrift for Jaegwon Kim*. T. Horgan, M. Sabates and D. Sosa (eds.). Cambridge. Mass.: MIT Press.

Block, N. and R. Stalnaker (1999). Conceptual analysis, dualism and the explanatory gap. *Philosophical Review* 108: 1-46.

Braddon-Mitchell, D. & Jackson, F. (2007). *Philosophy of Mind and Cognition*. 2nd Edition. Oxford: Blackwell.

Chalmers, D. (2000). What is a neural correlate of consciousness? *Neural Correlates of Consciousness: Empirical and Conceptual Issues*. T. Metzinger. Cambridge, Mass., MIT Press. Pp. 17-40.

Chalmers, D. (2002). Consciousness and its place in nature. *Philosophy of Mind*. D. J. Chalmers. Oxford, Oxford University Press. Pp. 247-272.

Chalmers, D. (2006). Phenomenal concepts and the explanatory gap. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. T. Alter and S. Walter. Oxford, Oxford University Press. Pp. 167-194.

Churchland, P. (1985). Reduction, Qualia, and the Direct Introspection of Brain States. *Journal of Philosophy* 82: 8-28

Hill, C. (1991). *Sensations: A Defense of Type Materialism*. Cambridge, Cambridge University Press.

Hill, C. and B. McLaughlin (1999). There are fewer things in reality than are dreamt of in Chalmers' philosophy. *Philosophy and Phenomenological Research* 59: 445-454.

Hohwy, J. (2005). Explanation and Two Conceptions of the Physical. *Erkenntnis* 62(1): 71-89.

Hohwy, J. and C. D. Frith (2004). Can neuroscience explain consciousness? *Journal of Consciousness Studies* 11(7-8): 180-198.

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

Jackson, F. (1986). What Mary didn't know. *Journal of Philosophy* 83: 291-295.

Kim, J. (1998). *Mind in a Physical World*. Cambridge, Mass., MIT Press.

Kim, J. (2005). *Physicalism; or Something Near Enough*. Princeton, N. J., Princeton University Press.

Kim, J. (2008). Reduction and Reductive Explanation: Is One Possible without the Other? *Being Reduced*. J. Hohwy and J. Kallestrup (eds.). Oxford, Oxford University Press. Pp. 93-114.

Kripke, S. (1972/1980). *Naming and Necessity*. Cambridge, Mass., Harvard University Press.

Levine, J. (2006). Phenomenal concepts and the materialist constraint. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. T. Alter and S. Walter. Oxford, Oxford University Press. Pp. 145-166.

Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy* 74(4): 549-567.

Lipton, P. (2004). *Inference to the Best Explanation*. 2nd Edition. London, Routledge.

Loar, B. (1997). Phenomenal states (second version). *The Nature of Consciousness: Philosophical Debates*. N. Block, O. Flanagan and G. Güzeldere. Cambridge, Mass., MIT Press. Pp. 597-616.

Melnyk, A. (2003). *A Physicalist Manifesto*. Cambridge, Cambridge University Press.

Papineau, D. (1998). Mind the Gap. *Nous* 32(S12): 373-388.

Papineau, D. (2002). *Thinking about Consciousness*. Oxford, Oxford University Press.

Polger, T. (2004). *Natural Minds*. Cambridge, Mass., MIT Press.

Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *The Philosophical Review* 117(2): 245.

Shoemaker, Sydney. (2007). *Physical Realization*. Oxford: Oxford University Press.

Smart, J. J. C. (1959). Sensations and Brain Processes. *Philosophical Review* 68: 141-156.

Stoljar, D. (2005). Physicalism and Phenomenal Concepts. *Mind and Language* 20(5): 469-494.

Tye, M. (1995) *Ten Problems of Consciousness*, MIT Press.

van Gulick, R. (1993). Understanding the phenomenal mind: Are we all just armadillos? *The Nature of Consciousness: Philosophical Debates*. N. Block, O. Flanagan and G. Güzeldere. Cambridge, Mass, MIT Press. Pp. 435-442.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science* 69: S366-S377.

Woodward, J. (2003). *Making Things Happen*. New York, Oxford University Press.