David Papineau

What Exactly is the Explanatory Gap?

1 Introduction

It is widely agreed among contemporary philosophers of mind that science leaves us with an 'explanatory gap'—that even after we know everything that science can tell us about the conscious mind and the brain, their relationship still remains mysterious.

I think that this agreed view is quite mistaken.

The feeling of an 'explanatory gap' arises only because we cannot stop ourselves thinking about the mind-brain relation in a dualist way.

So the 'explanatory gap' doesn't represent some unfinished business which remains even after we have learned everything that science can teach us. Rather we think that there is an explanatory gap only because we haven't yet properly embraced the findings of science.

Even though modern science shows us that dualism is false, we find it very difficult to absorb this lesson. This alone makes us think that there is something mysterious about the mind-brain relation.

In this paper I shall first introduce the idea of an 'explanatory gap' (section 2), and then rehearse the standard philosophical account of why it arises (section 3). I shall then criticize this standard account (section 4) and defend my own view that the impression of a gap is due to nothing but an intuitive conviction that dualism is true and materialism false (section 5). I conclude by explaining how we should respond to this dualist intuition (section 6) and by considering some possible explanations for its persistence even in the face of contrary evidence (section 7).

2 The Explanatory Gap

Let us take it that materialism holds that there are identities between conscious kinds and material kinds.[1] So, for example, it might turn out that pain is identical with the firing of C-fibres, or that seeing something as red is identical with such-and-such an activity in the V4 area of the visual cortex.

And let us also take it, in line with contemporary materialist orthodoxy, that any such identities will be a posteriori. It is a matter for scientific investigation, rather than conceptual reflection, to ascertain

---

[1] Putting it like this might make it seem I am committing materialists to 'type-identity' and ignoring the possibility of 'non-reductive physicalism'. But this is not my intention. Suppose, in line with non-reductive physicalism, that a conscious kind like pain metaphysically supervenes on physical properties without being identifiable with any of them—it is 'variably realized', by C-fibres in humans, but by different physical properties in octopuses, extra-terrestrials, androids, and so on. Then I take it that the relevant variably realized conscious kind must have a nature which explains why it so supervenes on physical properties. For example, it may be identical with the higher-order property of having some property which plays a certain causal role, or it may be a determinable property with a range of physical determinates, or it may be a disjunctive property with a range of physical disjuncts, and so on. And so, in any such case, we can identify the conscious property with the generic higher-order/determinable/disjunctive property, even if we can't identify it with any strictly physical property. (I shall illustrate some of the arguments that follow by assuming such identities as that of pain with the 'firing of C-fibres'. But this is an expository convenience. In truth, I am neutral about which kinds of material states—generic or more strictly physical—mental states are identical with.)

whether pain is the firing of C-fibres, or seeing something as red is activity in V4.[2] In this respect, mind-brain identities that we discover will be akin to such paradigmatic a posteriori scientific identities as that water is H_2O, or that temperature is mean kinetic energy.

And, finally, let us take it that there is good scientific evidence for the existence of such mind-brain identities. In some cases we will have direct evidence for the co-occurrence of specific conscious and material kinds. And even when we lack such direct evidence, there is more general evidence, in the form of the 'causal closure' of the physical realm, which argues that each conscious kind must be identical to some material kind (even if we don't know which), otherwise conscious causes couldn't have the physical effects they manifestly do have. (Cf. Papineau 2002 Ch 1 and Appendix.)

Even so, as Joseph Levine pointed out nearly thirty years ago, mind-brain identities strike us quite differently from the supposedly analogous scientific identities. (Levine 1983.)

Suppose that we really did have good evidence that pain is one and the same as the firing of C-fibres. Wouldn't we still want to know why there is pain when C-fibres are firing? Why do C-fibres firings feel like that, rather than like something else, or like nothing at all?

No analogous questions press in on us in the scientific cases. After we find out that water is H_2O, we don't feel that we still need to know why there is water when there is H_2O, or why H_2O manifests itself as water rather than in some other way. Water is H_2O—and that's that.

Levine coined the phrase 'the explanatory gap' for the impression that there is something left unexplained by mind-brain identities.

3 The Conventional Philosophical Story

It is very widely assumed by contemporary philosophers that the explanatory gap arises because we cannot derive the facts about conscious states a priori from the physical facts, in a way that we can derive facts about water, and facts about temperature, and so on, from the physical facts. Moreover, it is widely assumed that this contrast in derivability arises because of the different ways in which we conceive conscious states, on the one hand, and natural kinds like water and temperature, on the other.

Consider once more the identity of water and H_2O. This is of course an a posteriori matter. Merely possessing the concepts water and H_2O does not ensure that you know this identity. Still, suppose that you had a posteriori knowledge of all the physical facts, including all physical facts about the behaviour of H_2O in different circumstances. In that case, it is arguable that you will be able to move a priori from this physical knowledge to the identification of water with H_2O.

The idea here is that our concepts of natural kinds like water are descriptive concepts. We think of water as that liquid, whatever it is, that is tasteless, odourless, colourless, flows in rivers, expands on freezing, and so on. This then enables us to discern where the water is, so to speak, in a full physical description of the world. We just check through that description to ascertain which physical liquid satisfies the descriptions that a priori constitute our concept of water.

But now consider the identity pain and the firing of C-fibres, or any other identity between a conscious kind and a material kind. It seems clear that no amount of physical knowledge will allow you to derive the pain facts a priori from the physical facts. Scrutinize the physical facts about the behaviour of C-fibres as much as you like, and they will not tell you that it is painful to have your C-fibres firing.

---

[2] Similarly it is a matter for scientific investigation whether specific conscious states are identical with specific higher-order/determinable/disjunctive states.

The barrier here is that our primary concepts of conscious kinds like pain or seeing something as red are not descriptive concepts. We think of such kinds directly, in terms of what they are like, as it were, rather than as entities that play some descriptive role. You don't need to know about the causal role or other characteristics of pain or of seeing something as red in order to be able to think about these experiences. If you have had these experiences yourself, you will be able to think of them directly.

Because of this, there is no a priori route from the physical facts to the identification of pains with C-fibre firing. Since we don't think of pain as the state that plays such-and-such a behavioural role, we won't be able to appeal to any such a priori descriptive knowledge to figure out which physical state plays the pain role.

It was David Chalmers and Frank Jackson who first focused philosophical attention on our special way of thinking about conscious states. Jackson's 'knowledge argument' showed that there is a way about thinking about the experience of which is a priori distinct from any descriptive knowledge of the experience and which normally depends on previously having had that experience oneself (Jackson 1983). And Chalmers coined the term 'phenomenal concept' to refer to this distinctive way of thinking about conscious states (Chalmers 1996).

The past two decades have seen an intense debate about the existence and precise nature of phenomenal concepts. But we can by-pass that here. The only point that matters for present purposes is that there we can think of conscious states in a manner which does not allow us to derive facts about those states a priori from the physical facts. This much is very widely accepted among contemporary philosophers of mind, and I shall mean no more than this when I talk of 'phenomenal concepts' in what follows.

Chalmers and Jackson have argued that our inability to derive facts about consciousness a priori from the physical facts is inconsistent with a materialist view of consciousness. However, this argument hinges on contentious semantic and metaphysical premises and is not widely accepted. Most contemporary philosophers are happy to allow that mind-brain identities can be true even if there are no descriptive concepts of conscious kinds that might allow a priori derivations of such identities from the physical facts.

Even so, while few contemporary philosophers agree with Jackson and Chalmers that the non-derivability of conscious facts from physical facts discredits materialism, most of them do suppose that this non-derivability accounts for the appearance of an 'explanatory gap'. That is, they hold that the reason pain = C-fibres firing seems to leave something unexplained, where water = H_2O does not, is that the former identity is not derivable from the basic physical facts, whereas the latter identity is so derivable.

4 The Failings of the Conventional Account

In my view, this conventional account of the 'explanatory gap' is quite mistaken. Our contrasting reactions to mind-brain and scientific identities have nothing to do with differences in derivability from the physical facts.

Levine is of course quite right to observe that we react quite differently to mind-brain and scientific identities, and that our reaction to the mind-brain identities involves a feeling that something is left unexplained in these cases.

But this is not a feeling that we are left with after we have accepted the mind-brain identities, a feeling somehow occasioned by our awareness that we cannot derive these identities from the physical facts in the way we can supposedly derive scientific identities. On the contrary, the feeling is simply due

to our resistance to the mind-brain identities themselves. We find it very difficult to believe that pains just are C-fibres firing, or any similar equation of conscious states with material ones.

It is simply this inability to free ourselves from dualist thinking that makes us think that something is unexplained in the mind-brain cases. To the extent that we view pains and other conscious states as metaphysically independent from brain states, we of course face obvious explanatory questions. Why do certain physical processes but not others have the special power of exuding some mysterious mind stuff? And why do they produce just those conscious feelings, and not others? These are real enough questions, but they are not questions somehow prompted by the a priori underivability of the relevant mind-brain identities. They arise from nothing more than a simple refusal to accept the identities in the first place.

There can be a convergence with Henry's arguments: dualism is the result of a way of thinking. Once that way is entrenched, dualism seems inescapable.

The remainder of this section will be devoted to the failings of the conventional view that the explanatory gap is a matter of a priori underivability. In the next section I shall offer support for my alternative thesis that it stems from an implicit commitment to dualism.

The first point to make about the conventional view is that there seem to be plenty of other identities, apart from mind-brain identities, which are not derivable from the physical facts, and yet which generate no feeling of an explanatory gap. This point is defended at length by Ned Block and Robert Stalnaker (1999) and has been responded to by advocates of the conventional view (Chalmers and Jackson 2001). I do not intend to add to the intricacies of this debate here. For what it is worth, it seems obvious to me that identities involving proper names (Saul Kripke is the greatest living philosopher) or indexicals (today is Tuesday) cannot be derived a priori from the physical facts, given that proper names and indexicals lack any descriptive meanings. Yet we do not feel that there is anything left unexplained by identities involving proper names or indexicals. Moreover, I strongly doubt that even such paradigm scientific identities as water = H_2O can in fact be derived a priori from the physical facts either (since I doubt that terms like water really have descriptive meanings of the right kind). Yet here too difficulties about a priori derivability engender no impression that there is anything left unexplained by the relevant identities.

Still, as I said, I do not want to dwell on these well-trodden issues. Instead I want to question the underlying idea that a priori derivable identities have some power to explain things that are left unexplained by identities that are accepted on ordinary empirical grounds. Despite the widespread assumption that a priori derivability offers some explanatory advantage, it is very obscure how this is supposed to work. In particular, it is hard to see what exactly is supposed to get explained in such cases. Precisely what facts can be explained when identities are derived a priori, but not when they are accepted on ordinary empirical grounds? It is surprisingly difficult to answer this challenge. Let me consider various possible candidates in turn.

Explaining Identities? An obvious first thought is that it is the identities themselves that gets explained. If the physical facts plus the descriptive content of the concept water together imply a priori that water = H_2O, won't this explain why water = H_2O? By contrast, if we have to accept an identity on ordinary empirical grounds—pain = C-fibres firing, say—then won't we lack an explanation of why the identity holds?

But this first thought does not stand up. Identities need no explanation. Identities are necessary. They could not have been otherwise. So they are not the kind of fact that calls for explanation. (To repeat a familiar story, when you discover that Mark Twain = Samuel Clemens, you might reasonably ask why he had two names, or why nobody told you before. But it would make no sense to ask—why was Mark Twain = Samuel Clemens?)

Explaining Belief in Identities? Perhaps the thought is that without a priori derivability we can't explain why we should believe mind-brain identities. (The derivability gap is sometimes termed the 'epistemic gap'.) Frank Jackson has explicitly maintained that materialists would have no basis for believing that any everyday kinds are identical with material kinds if they could not derive these

identities a priori from the physical facts (Jackson 2003 254-5). But it is hard to see why he thinks this. Identities can be evidenced more directly in a number of ways. We might simply observe that the two kinds at issue co-occur. Or we might note that the physical kind in question shares some behaviour with the everyday kind. For example, it would be pretty good evidence that H_2O is water that it expands on freezing in just the way that H_2O does.

You might think that this latter kind of appeal to the behavioural role of the everyday kind is just the kind of thing that Jackson has in mind—we note that the physical kind plays the same role as the everyday kind. However, contra Jackson, there is no requirement here that the relevant behaviour of the everyday kind be a priori built into our concept of the kind. The freezing behaviour of H_2O would be just as evidentially significant if our knowledge of water's freezing behaviour derived from a posteriori observation rather than from our concept of water (as it surely did). Similarly, it may be crucial to the identification of pain with the firing of C-fibres to know that C-fibre firings are caused by bodily damage and in turn cause avoidance behaviour—but the identification will be just as well evidenced if our knowledge of pain's causal profile is a posteriori and no part of the relevant phenomenal concept.

Explaining Realization? Perhaps the explanatory significance of a priori derivability relates to realization rather than identity.

I said earlier (footnote 1) that I didn't intend my focus on identity claims to rule out non-reductive physicalism—a mental kind can be a posteriori identical with some generic higher-order/determinable/disjunctive kind while at the same time being variably realized at the strictly physical level. In such a case, there is room to ask why the relevant strictly physical states realize the generic kind. Perhaps we need a priori derivations to answer questions of this kind.

Suppose that some mental state is realized by, but not identical to, some strictly physical state. We might then want to explain why this strictly physical state realizes the mental state—why C-fibre firing realizes pain, say, or such-and-such activity in V4 realizes seeing something as red. And answering these questions might seem to require that we have some a priori conception of pain or seeing something as red a kind of generic property that will be determined by strictly physical states of a certain kind.

But this line of though does not stand up. I agree that if mental states are realized by strictly physical states we will want to understand why this is so, and I agree that in order to explain this we need to know what kind of property constitutes the generic realized mental state. However, this latter knowledge need not be conceptually derived—the explanation will work just as well if it is an a posteriori discovery.

So, for example, let us suppose that C-fibre firing realizes pain in humans without being identical to it. To explain why this is so, we will need to know about the generic nature of pain—perhaps that it is the higher order state of having some state that is caused by bodily damage and in turn causes avoidance behaviour. This knowledge will then enable us to explain why C-fibre firing realizes pain, by showing how it plays the causal role constitutive of this higher-order state. But nothing in this requires that this identification of pain with the higher-order state is built into our concept of pain. The explanation will work just as well if we use empirical evidence about the causes and effects of pain to establish which higher-order state pain is identical to.

Explaining Behaviour? Perhaps the point is that a priori derivability allows us to explain the characteristic behaviour of everyday kinds. So, for example, once we know that water is H_2O, we are in a position to explain why it displays such defining characteristics as odourlessness, colourlessness, and tastelessness—by showing that H_2O itself has just these characteristics.

But this is no good either. It is true that we can often explain the characteristic behaviour of some everyday kind in the light of knowledge of its physical nature. But there is nothing in such

explanations which requires that our concept of the everyday kind is defined in terms of that characteristic behaviour and that our knowledge of its physical nature is somehow derived a priori from this plus the physical facts.

Thus suppose once more that we start with a phenomenal concept of pain. As before, we might use ordinary a posteriori evidence to establish that it is identical with C-fibre firings. We might also note a posteriori that pain is characteristically caused by bodily damage and gives rise to avoidance behaviour. And then we will be in a position to explain the latter in terms of the former. Given the identity of pain with C-fibre firings, we can use the facts that C-fibre firings are themselves caused by bodily damage and gives rise to avoidance behaviour to explain why pain has those characteristics. Once more, a posteriori mental-material identities seem no less explanatory than a priori derivable identities.

5 The Real Gap

Whichever way we turn it, it is hard to discern any fact that is left unexplained by mental-material identities accepted on a posteriori grounds. There seems no good basis for the idea that a priori derivability yields some extra explanatory power.

I conclude that, if only we were able fully to persuade ourselves of mind-brain identities, we would see that they leave nothing unexplained. The problem is that we find it very hard to believe such identities. We all experience an intuitive resistance to identifications of phenomenal kinds with material kinds. At an intuitive level, we are all implicit dualists. Is that true for pre-galilean people?

And of course, as I explained at the beginning of the last section, such intuitive dualist thoughts will themselves generate a kind of explanatory gap. They will make us hanker to know why certain physical processes exude extra conscious states, and in particular why they exude just those feelings, and not others.

Let me now back up my claim that we are all in the grip of an intuition of dualism. Some of those who profess materialism at a theoretical level may be surprised to be told that they are closet dualists. But it is not hard to back up this diagnosis. Consider the terminology normally used to discuss the relation between mind and brain. Brain processes are standardly said to 'generate', or 'yield', or 'cause', or 'give rise to' conscious states. These expressions are common currency among many thinkers who will insist that they are no dualists. But the phraseology gives the lie to their denial. Fire 'generates', 'causes', 'yields' or 'gives rise to' smoke. But H_2O doesn't 'generate', 'cause', 'yield' or 'give rise to' water. It is water. Then point should be clear. To speak of brain processes as 'generating' conscious states, and so on, only makes sense if you are implicitly thinking of the consciousness as ontologically additional to the brain states.

Here is a more theoretical argument to the same conclusion. Consider a zombie, that is, a being who shares all your physical states but has no conscious mental life. Does this being strike you as initially possible? It take it that for nearly everybody the answer is 'yes'. There doesn't seem anything immediately metaphysically incoherent about such a being. Of course, philosophically sophisticated materialists will know that they are committed to denying that zombies really are possible. If conscious states are one and the same as brain states, then you can't possibly have one without the other. Still, it is not this reflective denial of the zombie possibility that I want to focus on here, but the initial intuition that zombies are possible, an intuition that I take to be present in even reflective materialists.

It is surprising that zombies strike materialists as even initially possible. Take other cases where we take ourselves to know some identity—that Cicero = Tully, say. And now ask yourself whether it is possible to have Cicero without Tully, or vice versa. The natural reaction is that there is simply no such possibility. If there is really just one person in question, what are we supposed to be supposing? That Cicero might exist without himself? That is blatantly incoherent. But then it is puzzling that a

materialist who believes, say, that pain is one and the same as the firing of C-fibres should be open to the zombie possibility that a being might have C-fibre firings but feel no pain. Isn't this equally to suppose that something might exist without itself? Why doesn't this strike materialists as blatantly incoherent too?

I infer from this that materialists don't fully believe their materialism. If they did, zombies would strike them as absurd to them as Cicero without Tully. The fact that they are open to the possibility of a being with C-fibre firings but no pain can only mean that they don't really believe the two states are identical.

You might object that the analogy is not a good one. In the case of Cicero = Tully and many other everyday identities, there is the possibility of reading the terms involved descriptively, which will then point us to some genuine possibilities. So, for instance, if we understand 'Cicero' as 'the greatest Roman orator' and 'Tully' as 'the greatest Roman statesman', then we can comprehend the idea of Cicero without Tully as alluding to the perfectly genuine possibility that the greatest Roman orator might not have been the greatest Roman statesman.

But this only drives the point home. There is indeed a disanalogy between Cicero = Tully and similar identities, on the one hand, and pain = C-fibre firings, on the other. Where terms like 'Cicero' can be understood descriptively, there is no corresponding descriptive reading of phenomenal terms like 'pain'. Understood phenomenally, the term 'pain' picks out it referent directly, in terms of what it is like, and not via any contingent description. But this only makes it all the more anomalous that materialists should have any space for the though that zombies are possible. The disanalogy should make it easier to posit Cicero without Tully than pains without C-fibres firing, because of the possibility of a descriptive reading. But in truth it is harder. Despite the putative availability of a descriptive reading, the natural reaction to the hypothesis of Cicero without Tully is incomprehension. Yet even without any room for a descriptive reading, materialists don't find any initial difficulty with the idea of a zombie. Since there is no question of understanding this latter possibility descriptively, I conclude once more that materialists don't fully believe their materialism.[3]

6 Living with the Intuition of Dualism

In arguing that all of us, professed materialists included, are in the grip of an anti-materialist intuition, I am not intending to raise a problem for materialism. The points made so far in this paper argue that materialism is a well-evidenced, cogent position which when properly understood leaves nothing about the relation between mind and matter unexplained. Against this background, it is no argument against materialism that people find it hard to believe. Many truths are hard to believe.

If the intuitive implausibility of materialism does raise a problem, it is a problem for materialists, not for materialism. Materialists need to recognize the insidious grip of dualist intuitions and adjust their thinking accordingly.

Some readers might feel that the intuition of dualism is at least some argument against materialism. Maybe it is not conclusive. But don't materialists owe us some explanation of this intuition? Don't they need to explain it away, by showing why it arises even thought it is false? And until we are given such an explanation, shouldn't we regard the dualist intuition as prima facie evidence against materialism?

I do not concede even this much. Of course, the dualist intuition is noteworthy, and we would like to explain it if we can. And accordingly in the next section I shall briefly consider the prospects for such

The analysis of the Galilean decision (Henry) is a possible explanation

[3] This argument is of course inspired by the final sections of Kripke's Naming and Necessity 1980. In a previous paper I defended the exegetical view that this was Kripke's own argument (Papineau 2007). I am no longer entirely sure of this.

an explanation. But it is not to be taken for granted that, in the absence of such an explanation, the dualist intuition supports dualism rather than materialism.

The issue here is whether the psychological fact that people find dualism intuitively plausible is evidence in favour of dualism.[4] Well, this would be good evidence if dualism offered a better explanation of this psychological fact than materialism. (I take a fact to support a theory to the extent that the theory makes the fact more probable than its competitors do.) But it is by no means to be taken for granted that dualism can offer a better explanation for the dualist intuition.

As a general rule, the truth of p is the most obvious explanation for people believing that p, and to this extent it is reasonable to take their belief in p as evidence for p. But the rule is easily defeated. There are cases where we can see that the belief in p can't be due to the truth of p, and in such cases the fact that people believe p ceases to provide any support for p itself.

The intuition of dualism is a case in point. It is arguable that dualism requires epiphenomenalism. To suppose otherwise requires denying the causal closure of the physical, a step which few contemporary dualists are prepared to take. But, if dualism requires epiphenomenalism, then it is hard to see how dualism can possibly explain any beliefs, let alone beliefs in the truth of dualism.

We can take it that beliefs themselves are denizens of the material world, manifesting themselves in verbal and other behaviour, as for example when people manifest their belief in dualism by saying there is an explanatory gap between matter and mind. So, if non-material dualist phenomena have no influence on the material world, as required by epiphenomenalism, then they cannot cause beliefs, and in particular cannot cause beliefs in dualism. Whatever the explanation of people's conviction that dualism is true, it cannot be the influence of dualist phenomena on their thinking, because there can be no such influence.

This means that, even if dualism is true, the cause of dualist beliefs must be material influences on people's thinking, not dualism itself. So dualism makes belief in dualism no more likely than materialism does. Conversely, then, the fact that people believe dualism gives us no reason to suppose that dualism is true. They would be just as likely to believe dualism if it were false.

So the intuition of dualism is no evidence at all for the truth of dualism. Still, what are materialists supposed to do with the intuition of dualism? It is awkward, to say the least, to find yourself continually judging at an intuitive level something that you are theoretically committed to denying.

Well, one possibility is that the intuition of dualism will fade away as materialism wins adherents. However, as I shall explain in the next section, I don't think that this is likely. I suspect that the intuition derives from some deep-seated feature of our cognitive architecture, and will continue to press on us even after the arguments for materialism becomes orthodox and familiar.

If this right, then materialists will just have to live with the intuition. In a sense, they will be stuck with contradictory beliefs. At a theoretical level, they will recognize the strength of the evidence for materialism and be committed to its truth. But at the same time they will continue to experience an intuitive conviction that materialism is false. They will be able to discount this conviction in their theoretical discourse, but it will nevertheless remain present at an intuitive level.

This is an odd set-up, but by no means unique. There are many other cases where we cannot shake off an intuitive belief that we know to be false at a theoretical level. At a theoretical level, I know that

---

[4] Note that I am here reading 'the intuition of dualism' as referring to a psychological attitude, not the propositional content of that attitude. Of course the content of the intuition would support dualism—if it were granted. However the issue at hand is the different one of whether the psychological fact that people intuitively believe dualism is evidence for dualism.

the earth is moving, but intuitively I feel that I am standing on firm and unmoving ground. At a theoretical level, I am convinced that there is no moving present and the B-series description of reality is complete, but at an intuitive level I can't stop myself thinking that I am moving through time. At a theoretical level, I am persuaded that reality splits into independent branches whenever a quantum chance is actualized, but at an intuitive level I can't get rid of the belief that either I will develop cancer or I won't. And so on.

7 Possible Explanations

If dualism is false, then why are we all, materialists included, in the grip of an intuition that it is true? In this section I shall briefly consider some possible explanations for this intuition.

Ingrained Culture

One possibility is simply that many of us are brought up as dualists. Western culture has long taken dualism for granted. Dualism plays a central role in Christian theology, and until recently was also supported by mainstream scientific thought (Papineau 2002 Appendix). Maybe we need look no further for an explanation of why dualist thinking comes so naturally to all of us.

However, this hypothesis has the implication that, if our culture were to embrace materialism whole-heartedly, then dualist intuitions would dissolve away. Some philosophers are happy with this implication. Thus Richard Rorty (1979) has contended that a community that grew up with materialism would view the mind-brain relation as quite unproblematic. And Stephen Yablo asks: 'Am I the only one who feels the intuition of zombies to be vulnerable in this way?' (2000 119).

I am not convinced. I suspect that there is something more structural pushing us towards dualism, some feature of our cognitive architecture that forces the intuition of dualism on us. If that is right, then the intuition won't be removed just by a simple change of culture.

Natural Born Dualists

The Yale psychologist Paul Bloom agrees that our dualist inclinations are imposed on us by our cognitive architecture. In Descartes' Baby (2004) he argues that human infants automatically develop a dualist view of the world. As he sees it, this is an upshot of our having two distinct cognitive systems for thinking about mental and material processes respectively. On the one hand we have the 'mindreading' module which leads us to attribute mental states to intentional agents; on the other we have the 'folk physics' module which we use to reason about the material world. Any given phenomenon will activate one or the other module but not both—which according to Bloom is why we view mind and matter as two incommensurable realms.

This is an intriguing suggestion with genuine explanatory power. It certainly accounts nicely for the extreme ease with which we comprehend the possibility of people 'swapping bodies', as in many familiar fictions and philosophical thought experiments.

However, I am not sure that it gets to the bottom of the intuitive grip that dualism exerts on us. The problem is that Bloom's story seems to overgeneralize.

Our 'mindreading' modules allow us attribute, not just conscious states with a phenomenology, but intentional states in general. For example, I might judge that you believe that Jack stole the tarts, without necessarily supposing that this belief has any active phenomenology in you.

So, if Bloom's story were right, we ought to have strong dualist intuitions even about such implicit mental states which lack any associated 'what-it's-likeness'. However, I take it that we don't. We don't feel that your implicit belief that Jack stole the tarts must somehow be non-physically realized—precisely because there isn't any phenomenology to this state. It looks as if the intuition of dualism

hinges on the way we think specifically about phenomenological states, rather than about mental states in general.

The Antipathetic Fallacy

States with a phenomenology can be thought about using phenomenal concepts. I said earlier that I would assume nothing about phenomenal concepts except that they are a priori distinct from material ways of thinking about conscious states. But let me now add in one further familiar thought about phenomenal concepts—their exercise is often accompanied by an actual or imagined instance of the phenomenal state being referred to. For example, when I think phenomenally about the experience of seeing something red, I will often either actually be seeing something red or be recreating this experience in imagination.[5] Hum..

I have long argued that this feature of phenomenal concepts can confuse us into thinking that mind-matter identity claims cannot be true. For example, take the claim that:

(1) The experience of seeing something as red = such-and-such neural activity in V4.

When we entertain this identity claim, our phenomenal reference to the experience on the left-hand side will often be accompanied by an actual or imagined experience of seeing something red. By contrast, our reference to neural activity in V4 on the right-hand side will involve no such colour experience. This can lead us to think that the right-hand side of the identity 'leaves out' the experience referred to on the left-hand side. And so we naturally conclude that the left-hand side must be referring to something additional to the physical state referred to on the right-hand side—that is, that the experience of seeing something as red is something over and above neural activity in V4.

Interesting argument

Now, of course this line of thought is a fallacy, engendered by a sort of use-mention confusion. That we do not activate or 'use' the experience on the right-hand side does not mean that we are not there referring to ('mentioning') to that selfsame experience—it is certainly not in general required that we can only refer to something if it is itself somehow activated in the act of referring. But for all that it has always seemed to me a very seductive fallacy. (In Papineau 1993 I termed it 'the antipathetic fallacy'.)

Recently Pär Sundström (2008) has queried this account of the source of our dualist intuitions. He argues that it predicts dualist intuitions in cases where there are none. For example, consider the claim

(2) The experience of seeing something as red = John's most salient current experience.

This seems analogous to the earlier identity in the relevant respects. Contemplation of the left-hand side is likely to be accompanied by the experience in question, but not so for the right-hand side. So if the antipathetic fallacy were the right diagnosis of the dualist intuition, the intuition ought to arise here too. But, as Sundström observes, it is not obvious that it does. We don't feel that the experience of seeing something as red must be something distinct from John's most salient current experience.

Sundström's point is well-taken, but I am not sure it is conclusive. Note that the antipathetic fallacy is supposed to occur with identities that do activate the experience on the left-hand side but don't on the right. In order for Sundström to have a clear counter-example to the antipathetic fallacy explanation he needs to be sure that his identity claim satisfies these requirements. Perhaps this gives the

---

[5] On some accounts of phenomenal concepts, this accompaniment is essential to the exercise of phenomenal concepts, while on others it is merely a frequent corollary. This difference will not matter in the present context.

antipathetic fallacy explanation some room for manoeuvre. Maybe in cases like Sundström's we tend surreptitiously to activate the experience on the right-hand side when we refer to John's current experience—we fill out the reference by imagining the experience of seeing something as red. Alternatively, maybe in his cases we tend not to activate the experience on the left-hand side. (My current view is that such activation is by no means essential to every deployment of phenomenal concepts—see Papineau 2008.) Either of these possibilities could explain, consistently with the general lines of the antipathetic fallacy suggestion, why there is no intuition of dualism in cases like (2) as opposed to (1). It would be useful to have further empirical investigation of this issue.

Aspirations of Transparency

Consider the following line of thought:

"When we think of conscious states phenomenally, in terms of what they are like, we are in direct contact with them--the phenomenal states are right there before the mind, so to speak. Given this, all the essential properties of phenomenal states should be transparent to phenomenal thinking. However, phenomenal thinking does not reveal phenomenal states to be physical. So such states cannot be essentially identical to physical states."

Some philosophers defend something like this as a substantial argument against a materialist view of the mind. (Nida-Rümelin 1998, Goff forthcoming.) But the soundness of this argument is not the issue here. (Materialists will of course deny that phenomenal thinking has the assumed revelatory powers.) The present issue is rather whether the plausibility of this line of thought accounts for the general prevalence of dualist intuitions. Are people in general intuitively attracted to dualism because they assume that phenomenal introspection must reveal the essential nature of phenomenal states?

I am not sure. It might seem that this explanation of the dualist intuition assumes an overly sophisticated level of reflection about the mind-body issue on the part of ordinary thinkers. But in response it could be observed that people who haven't thought about the mind-body issue are unlikely to have any intuitions about dualism in the first place.

Merging the Files

Finally, let me consider a suggestion due to Andrew Melnyk.

One popular account of what happens when we accept any identity claim of the form a=b is that we 'merge the files'. Where we used to have two mental 'files' for 'a' and 'b' respectively, each containing different items of information, we reorganize our cognitive architecture so as to end up with one file containing the union of these items of information. (The alternative would be to keep two separate files, but copy any information entered in either across to the other. The merging option offers obvious efficiency gains.)

Melnyk's suggestion is that we aren't able to merge the files in the case of mind-brain identities. His thought is that phenomenal concepts and material concepts are realized in different parts of the brain, and that this is an obstacle to the normal merging of associated files. (Perhaps phenomenal concepts are associated with the sensory cortex, whereas material concepts are associated with linguistic areas of the brain.) So, when we are persuaded of a mind-brain identity claim, we are unable to carry out the merging operation that normally results from our accepting identities. This makes us feel that there is something amiss with mind-brain identities, and inclines us to dualist doubts.

This is an interesting suggestion, but there is an obvious danger that it too may predict more than it should. Presumably outward-directed perceptual concepts—concepts of colours, textures and other observable features of the non-mental world—are associated with the sensory cortex just as much as inward-directed phenomenal concepts. So, if Melnyk were right, shouldn't we be intuitively resistant to such perceptual-physical identities as redness = reflectance profile Ψ, or oiliness = surface structure

Ω, just as much as to pain = C-fibres firing? But it's not obvious, to say the least, that we resist these identities in the way we resist mind-brain identities.

Melnyk's response is that we are indeed resistant to perceptual-physical identities in just the way we are to phenomenal-physical ones, but that we tend not to regard this as generating an extra philosophical puzzle for materialism. This is because there is a natural tendency to think of perceptible properties like redness or oiliness as a combination of an external and subjective component—we can think of them, in Lockean style, as that external property, whatever it is, that causes subjective impressions of redness or oiliness in us. And this then allows us, when we reflect on perceptual-physical identities, to conclude that any puzzlement we feel about them is simply a reappearance of the familiar difficulty that arises with phenomenal-physical identities. That is, we tell ourselves that the external property involved can indeed be unproblematically identified with a reflectance profile, say, or a surface structure—and that any puzzlement we feel about perceptual-physical identities is therefore simply a reappearance of the familiar difficulty of understanding how subjective experiences of redness or oiliness can be identical to anything physical.

8 Conclusions

The last section shows that there are a number of possible explanations for the persistent intuition of dualism. All of them face some difficulties, but there is no reason to think that they are all insuperable.

In the end, of course, the evaluation of these suggestions is an empirical matter. We are addressing a psychological phenomenon—people in general experience a persistent intuition of dualism, even in the face of strong contrary evidence—and trying to figure out the cause of this psychological phenomenon. Resolving this issue properly will depend on empirical investigation, not further armchair speculation.

Note that there is no reason why there should be just one cause for the intuition of dualism. Perhaps a number of the suggestions considered in the last section contain some element of truth, and the intuition of dualism is a product of multiple factors pushing in the same direction.

It may seem unsatisfactory to leave the empirical explanation of the intuition unresolved. But my primary aim of this paper was not to explain why we experience this intuition, but to demonstrate that we do, and that this is the only reason why mind-brain identities strike us a leaving an 'explanatory gap'.

This alone is enough to set materialism on the right track. Once we recognize the existence of this intuition, we can see that the explanatory gap poses no argument against materialism itself. Nothing is left unexplained by mind-brain identities. The only difficulty facing materialism is psychological, not theoretical. We need to keep the theoretical arguments for materialism firmly in mind, and not allow ourselves to be distracted by unsupported contrary intuitions.

References

Block, N, and Stalnaker, R. 1999, 'Conceptual Analysis, Dualism, and the Explanatory Gap' Philosophical Review 108 1-46

Bloom, P. 2004 Descartes' Baby Heinemann

Chalmers, D 1996 The Conscious Mind Oxford University Press

Chalmers, D. and Jackson, F. 2001, 'Conceptual Analysis and Reductive Explanation' Philosophical Review 110 315-61

Goff, P. forthcoming ‘A Posteriori Physicalists Get our Phenomenal Concepts Wrong’ Australian Journal of Philosophy

Jackson, F. 1986. ‘What Mary Didn’t Know’, Journal of Philosophy 83 291-5

Kripke, S. 1980 Naming and Necessity Blackwell

Levine, J. 1983 ‘Materialism and Qualia:The Explanatory Gap’ Pacific Philosophical Quarterly 64 354-61

Nida-Rümelin, M. 1998 ‘On Belief About Experiences: An Epistemological Distinction Applied to the Knowledge Argument’ Philosophy and Phenomenological Research 58: 51–73

Papineau, D. 1993 Philosophical Naturalism Blackwell

Papineau, D. 2002 Thinking about Consciousness Oxford University Press

Papineau, D. 2007 ‘Kripke's Proof is Ad Hominem Not Two-Dimensional’ Philosophical Perspectives 21 475–494

Papineau, D. 2008 ‘Phenomenal and Perceptual Concepts’ in Alter, T. and Walter, S. eds Phenomenal Concepts and Phenomenal Knowledge Oxford: Oxford University Press

Rorty, R. 1979 Philosophy and the Mirror of Nature Princeton University Press

Strevens, M. 2009 Depth Harvard University Press

Sundström, P. 2008 ‘Is the mystery an illusion? Papineau on the problem of consciousness’ Synthese 163 133-143

Yablo, S. 2000 ‘Textbook Kripkeanism and the Open Texture of Concepts’ Pacific Philosophical Quarterly 81 98-122