CHAPTER 18

A POSTERIORI PHYSICALISM

Type-B Materialism and the Explanatory Gap

JOSEPH LEVINE

18.1 Introduction

Chalmers (2002) famously categorizes versions of materialism into various 'types'. For our purposes the only distinction that matters is between 'Type A' and 'Type B'. According to Type A materialism, (ideal) reflection on our mental and physical concepts reveals that the possibility of a being physically identical to a fully conscious human but without conscious experience can be ruled out a priori. Put another way, so-called 'zombies' are not even conceivable. According to Type B materialism, however, zombies are conceivable, but they are not metaphysically possible. The idea is that since the possibility of zombies cannot be ruled out a priori, they count as conceivable, but since in fact the mental supervenes on the physical,[1] zombies are not genuinely possible. I will not be concerned further with Type A materialism in this chapter, only mentioning it to contrast it with Type B, my focus here.

In the past (Levine 1998, 2001) I have argued that though the conceivability of zombies (along with several other thought experiments) manifests the existence of an explanatory gap between the physical and the mental (specifically consciousness—let that be understood in what follows unless otherwise stated), there was still good reason to adopt materialism, though the lack of explanatory import posed a problem. In other words, I claimed that Type B materialism had a serious epistemic defect—the explanatory

[1] Here, and in the rest of the chapter, by 'supervenience' I mean *metaphysical* supervenience, a stronger relation than *nomological* supervenience. The latter is consistent with dualism, as it only posits a lawful relation between the physical and the mental, and so allows for basic laws of nature that link the two.

gap—but it was still philosophically coherent and viable. In order to maintain Type B materialism in the face of zombie conceivability it is necessary to open another gap—between conceivability and possibility. Much of the debate over the viability of Type B materialism has concerned the connection between conceivability and possibility. I have mostly taken the side of Type B materialists on this particular question.

More recently (Levine 2014), I have come to doubt the viability of materialism in the face of the explanatory gap. As I see it, the inference to the rejection of materialism takes the form of an inference to the best explanation; the best explanation of the existence and persistence of the explanatory gap is that there is a genuine metaphysical gap. I now favor some version of emergentism. However, though I no longer defend materialism, I still believe that the conceivability argument does not by itself show materialism to be false. I still hold that the original defense against the conceivability argument, the one that relies upon opening a space between conceivability and possibility, works.

In this chapter I want to do three things: In Section 18.2 I will outline what I take to be the principal argument that the rejection of materialism follows rather directly from the conceivability of zombies and explain why I do not accept that argument. In Section 18.3 I will consider a related argument against materialism that might seem to follow from premises that I explicitly endorse, and show that it does not after all. In Section 18.4 I will then turn to an argument that some materialists have presented, and seems to follow from the considerations presented in Section 18.3, that the explanatory gap is not really such a problem for materialists after all. I will try to explain why my inference to the best explanation still goes through in the face of this argument.

18.2 The Conceivability Argument against Materialism

As there is a vast literature on the conceivability argument and its effect on Type B materialism, some of which I have contributed to already, I will keep my discussion here brief.[2] Here is how I see the dialectic. The anti-materialist argues that since zombies are conceivable—which means that any statement to the effect that a zombie exists cannot be ruled out a priori—this shows that they are possible. If zombies are possible, even if not actual, this undermines the commitment of materialists to the claim that phenomenal consciousness (metaphysically) supervenes on the physical facts. If there is a possible world in which the physical facts are as they are in the actual world, and yet a physical twin of a conscious person in the actual world is a zombie in this possible world, then phenomenal consciousness has been demonstrated not to supervene on the physical facts and so a core commitment of materialism is false.

[2] For starters, see the works by Chalmers and myself mentioned already, Block and Stalnaker (1999), Chalmers and Jackson (2001), and the collection of papers in Gendler and Hawthorne (2002).

The first move in defense of the supervenience thesis is to assimilate materialist identity claims (whether with neural states or functional states) to other theoretical identity claims discussed by Kripke (1980) and Putnam (1975). Just as it is conceivable that heat not be the motion of molecules, or water not be H_2O, and yet no one doubts the truth of these identity claims, so too the conceivable falsehood of materialist identity claims does not undermine their acceptability. Clearly, then, conceivability of some proposition P does not entail that P is possible.

Anti-materialists, however, have a comeback. There are a lot of different ways to characterize this response, but for my purposes the following works best. While undoubtedly we can coherently conceive of a situation in which water turns out not to be H_2O—for instance, it could have turned out to be XYZ—what does not seem conceivable is that a certain conditional, call it a 'supervenience conditional', should turn out to be false. We can form the relevant supervenience conditional as follows: Take all the basic, fundamental facts about a world—all the physical facts, and whatever else provides the supervenience base for all non-basic facts—and put them in the antecedent of the conditional, and then make the statement that water is H_2O, or that heat is the motion of molecules, the consequent. Let us symbolize this as $B \rightarrow M$ (where 'B' represents all the basic facts, and 'M' represents non-basic macro-facts). The claim, then, is that $B \rightarrow M$ is a priori. So if you were given a description of all the basic facts about a world, then from this description and your grasp of the relevant concepts, you could infer what water is, what heat is, etc.

There is nothing in the standard Kripke–Putnam story that explicitly conflicts with this claim about the a priori status of these supervenience conditionals, and so appeal to the conceivability of water not being H_2O does not undermine the claim. But once armed with this claim—let us call it the 'a priori entailment thesis'—it then follows that the conceivability of zombies presents a serious challenge to the claim that phenomenal consciousness supervenes on the physical facts. After all, put in all of the physical and functional facts you think are those on which consciousness must supervene, the resulting supervenience conditional will still not be a priori. Hence there is a principled difference between the zombie case and the cases of water and heat, and so the original conceivability argument for the possibility of a zombie remains intact.

There are two standard lines of reply by Type B materialists in the literature (and another two that are not so standard). I call one of the two standard lines the 'exceptionalist' line and the other the 'non-exceptionalist' one. The non-exceptionalist line is to deny the a priori entailment thesis in general, arguing that conditionals like '$B \rightarrow M$', even where the substituend for 'M' is about some non-mental, middle-sized object like water, are not a priori. The exceptionalist line allows that in most cases conditionals like '$B \rightarrow M$' are a priori, but argues that when it comes to phenomenal states the situation is different. The reason is that when we think about phenomenal states we employ so-called 'phenomenal concepts', and these concepts have special features that block a priori inferences from statements employing non-phenomenal concepts.

Now the 'phenomenal concept strategy', as it has been called, is employed not only to counter the conceivability argument but also to defend materialism against the

challenge of the explanatory gap.[3] Elsewhere (Levine 2007) I have criticized the phenomenal concept strategy as a response to the explanatory gap, and from that critique it follows that it does not serve to undermine the conceivability argument either. However, I do think the non-exceptionalist strategy works; that is, the Type B materialist should just reject the entire semantic framework on which the a priori entailment thesis is based.

As I see it, the principal division here is between a Fregean, or neo-Fregean theory that provides a substantive role for a mode of presentation in doing the meta-semantic work of connecting our concepts to their referents (whether objects, properties, or entities of some other ontological category), and a Russellian (or neo-Russellian) theory that does not. On the neo-Fregean theory part of what determines reference for a concept are a priori accessible conditions that must be satisfied by a candidate entity in order to qualify as the referent. Thus, when considering some supervenience conditional of the form 'B → M', one employs one's knowledge of the reference-determining conditions that constitute the modes of presentation of the relevant concepts in order to determine the truth of the conditional. On the neo-Russellian theory, there is no mode of presentation grasped by the subject that determines the referent. Rather, the meta-semantic conditions that connect concepts to their referents work 'behind the scenes', as it were, out of the ken of the subject. The standard conditions appealed to are causal chains and nomic dependencies. If there are no a priori accessible modes of presentation (for most atomic concepts, that is) then there is no semantic knowledge on which to draw to make supervenience conditionals a priori. If the inference from a complete description of the micro-physical facts to the statement that there is water in the glass is not a priori, then the fact that the inference from the micro-physical facts to the phenomenal facts is not a priori should not be a problem. After all, if water supervenes on the micro-physical despite the fact that the relevant supervenience conditional is not a priori, then phenomenal consciousness can supervene on the micro-physical as well.

As I said, I favor the non-exceptionalist response to the conceivability argument. On the neo-Russellian semantic framework the a priori entailment thesis never holds, so therefore there is no reason to expect it to hold in the case of the psycho-physical connection. A description of the entire supervenience base for some phenomenon—indeed, include the entire supervenience base for all non-basic phenomena—need not a priori entail a description of the relevant phenomenon despite necessitating it. Having said that, I will address one line of argument supporting the a priori entailment thesis in the next section. In the remainder of this section I want to briefly characterize the two non-standard replies to the conceivability argument/a priori entailment thesis mentioned above.

The first line of reply starts by accepting the neo-Fregean semantic framework.[4] Let us assume that we have a restricted stock of primitive concepts within which the modes of

[3] Again, there is an extensive literature on phenomenal concepts. The locus classicus is Loar (1997). See also the papers in Alter and Walter (2007).

[4] This line of argument is presented in Levine (2014).

presentation of all other concepts can be expressed. When the mode of presentation of a non-primitive concept is applied to the characterization of a possible world in terms of the primitive concepts (what Chalmers calls a 'scenario') the reference of that concept in that world is determined. (This applies only to 'worlds considered as actual', as the 2D theorists put it.) So, if you have a conditional of the form S → M, where 'S' stands for an exhaustive scenario description and 'M' is any statement involving non-primitive concepts, the truth value of the conditional, on this view, is a priori. Hence, it might seem, this view entails the a priori entailment thesis.

However, notice that earlier we used the schema 'B → M' to represent a supervenience conditional, and now we have used the schema 'S → M' to represent what I will call a 'scenario conditional'. What is the difference between 'S' and 'B'? 'S' stands for a complete description of a world in terms of our cognitive system's primitive concepts, while 'B' stands for a complete description of a world in terms of the concepts that represent its basic, or fundamental properties. The a priori entailment thesis concerns the supervenience conditionals, not the scenario conditionals. So if the neo-Fregean semantic framework is to lend support to the a priori entailment thesis, it must be that 'S' and 'B' amount to the same thing. But this would be so only if the concepts that are primitive for us represented the properties that are metaphysically basic for our world. But why think this is the case? In fact, it seems highly unlikely that evolved creatures like ourselves, living in the medium-sized and relatively slow-paced world that we do, would have as primitive concepts ones that represented the most fundamental properties of our world. What a coincidence that would be!

To make this a little less abstract and bring it more directly to bear on Type B materialism, let me put the point another way. What is supposed to be problematic for Type B materialism is the alleged conflict between the following two claims: (1) that a conditional of the form 'P → Q' (where 'P' stands for all the physical facts and 'Q' the phenomenal qualitative facts) is metaphysically necessary, and (2) that this conditional is a posteriori (which is why zombies are conceivable). But suppose, as Chalmers and Jackson (2001) themselves argue, that phenomenal concepts, the ones in which 'Q' is couched, are among our primitive concepts, and therefore among those that, in a sense, define all our non-primitive concepts. Then of course if you put the phenomenal concepts, together with whatever other primitive concepts there are (perhaps spatial concepts, the concept of cause, etc.), into the antecedent of a scenario conditional it will be a priori. But since the concepts within which 'Q' is couched are themselves primitive, if 'Q' is put in the consequent of a conditional—as in 'P → Q'—there is no reason to expect the resulting conditional to be a priori. Therefore, the fact that 'P → Q' is a posteriori does not provide any reason for doubting its necessity, and Type B materialism is off the hook.

There is one final line of reply that is more speculative than the others but still, I think, worth putting out there. Advocates of the conceivability argument, particularly Chalmers and Jackson, emphasize that when presented with a scenario conditional people have the ability to decide, or infer, what the referent of their various non-primitive concepts are, such as 'water' and 'heat'. One might object that even if people can normally do this (which involves a lot of idealization, but let that go for now), what they are doing may be

engaging in non-demonstrative, or empirical reasoning. If so, then there would be no basis for considering the relevant conditionals to be a priori. However, they retort that since all the empirical information has already been put into the antecedent of the conditional, any reasoning from the antecedent to the consequent must be a priori, as no empirical considerations remain (outside the characterization of the antecedent) to influence one's reasoning.

Suppose one admits that, for the reason just cited, that the inference from the antecedent to the consequent must be a priori. Still, one may not be forced to locate the source of its a priority in the semantic competence of the subject (or at least not solely). Perhaps our most fundamental rules of reasoning themselves—not derived from semantic analyses but from epistemic norms—are a priori. It might be that whenever one reasons from empirical premises to an empirical conclusion one is employing an a priori epistemic rule system. This rule system differs from demonstrative reasoning in that it is not truth-preserving; nevertheless, it may be a priori. I can imagine independent reasons for thinking this must be the case.[5] While I do not myself have a firm position on this question, it does seem to me a not implausible path for the Type B materialist to take.

18.3 Avoiding Brute Necessity

Above I described the non-exceptionalist reply to the conceivability argument. On this view, we reject the neo-Fregean assumptions underlying the argument, instead opting for a neo-Russellian semantics that more firmly severs the connection between the epistemic status of a proposition and its metaphysical status. However, it is just this severance of the connection between the epistemic and the metaphysical that some argue manifests the weakness in the Type B materialist's position.

The problem revolves around the notion of what Chalmers calls a 'strong necessity', and which others (including myself) have called a 'brute necessity'. Many philosophers find the idea that a metaphysical necessity could be a brute, inexplicable fact to be extremely implausible, if not downright incoherent. Here is how I see the argument for this view. The world is constituted by a set of facts, organized in such a way that some are basic and others are realized in, or constituted by, other facts. We can add laws of nature into the set of basic facts if we are non-Humeans about laws. When we explain various empirical phenomena, we do so by appeal to more basic phenomena, and this process eventually bottoms out in facts and laws that are brute and inexplicable. Why is the gravitational constant what it is and not some other value? Maybe there is an explanation in terms of deeper phenomena, but maybe there is not. It could just be that it is what it is and that is just a brute fact about our world.

[5] See Biggs and Wilson (2017) on the a priori nature of inference to the best explanation.

That there should be brute facts about our world is not surprising. After all, if there were not it would mean that the world really had to be the way it is, and then in some sense its structure should be both necessary and accessible a priori. But we know we need experience of the world—we have to get data from it—in order to figure out what it is like. Some aspects of our world, particularly involving its most fundamental structure, just happen to be the way they are, and they are what distinguish it from all of the other possible worlds there could have been. We might say that each possible world is distinguished from every other by the particular constellation of brute facts and laws that constitute it.

So that contingent propositions might be brutely, inexplicably true is not a problem. We expect there to be such propositions. But the point is that a contingent proposition's being brute makes sense precisely because a contingent proposition is specifically about the actual world. But could a necessary proposition be brute? Could some brute fact hold necessarily? Modal facts are not specifically about the actual world, but rather about the entire space of possible worlds. So could some fact about the entire space of possible worlds plausibly be brute? Many philosophers, myself included, think not. If bruteness is constitutively tied to what *just happens to be* the case—the realm of the empirical—then it cannot be a feature of what *must be* the case. Hence necessities require some explanation for their necessity.

What could explain a necessity? Well the traditional conception—that is, pre-Kripke—was that the three distinctions, analytic–synthetic, a priori–a posteriori, and necessary–contingent were all aligned. The notions of being true by virtue of meaning (and logic), being knowable independently of experience, and being true in all possible worlds went perfectly together. That some fact/proposition is necessary is then explicable by appeal to its logical/semantic/epistemic status. The explanation would go something like this. Why is P necessary? Well, because P is a priori. But how does the epistemic status underwrite the metaphysical status? What connects a proposition's being a priori to its being necessary? The answer lies in logic and semantics. It is because of its logical/conceptual structure that the proposition in question is a priori, as mere competence with the concept and with logic renders its truth self-evident. If it is self-evidently true, without need to consult the way the world actually is through experience, then it must be because it has to be that way. Hence we understand why it is necessary.

Now of course this picture has to be complicated in the light of the generally accepted examples of a posteriori necessities made famous by Kripke and Putnam. However, making the accommodation to fit these examples while keeping the constitutive connections just described is not that hard to do. All we need to add is a dimension of modality that allows for a contextual element to determine the referents of concepts with which we have an a priori competence. So, knowing that water is whatever in this world satisfies the 'watery' description is a priori, and when we add the a posteriori information that it is H_2O that in fact satisfies the watery description, we know that water is necessarily H_2O. The point is that it is our a priori accessible competence with the concept of water that bears the burden of explaining the necessity in question. Thus it is not a brute necessity.

Chalmers (2006) has put this relationship in characteristically vivid terms: he calls it the 'golden triangle', with modality, rationality, and semantics occupying the vertices of this triangle. He argues that our very notion of the modal is constitutively tied to our notion of rationality—thus ruling out brute necessity—and that semantics, or our semantic competence, is the link that connects the a priori and the necessary. The Type B materialist, by classifying zombies as impossible despite their conceivability, is guilty, then, of positing a brute necessity, and thus violating this constitutive link between the necessary and the a priori.

We can get to the same conclusion by a slightly different line of reasoning (though ultimately connected to the argument above). The fight over materialism concerning conscious experience has been fought, for the most part, over the question of supervenience. If zombies are possible, then phenomenal consciousness does not supervene on the physical facts and, since materialism entails supervenience, materialism would be false. The reason that supervenience has been the focus of the debate is that the anti-materialist arguments challenge it. However, most materialists admit that supervenience is too weak a relation to adequately capture the thesis of materialism. Instead, most opt for realization (or, nowadays, grounding, which, as far as I can tell, comes to the same thing).

Now one of the principal ways that realization is supposed to go beyond supervenience is that the former, as opposed to the latter, entails that there is an explanatory relation between the realization (and supervenience) base and the facts it realizes (and therefore, which supervene on it). As Jeffrey Poland (1994) puts it, if one set of facts realizes another, then there must exist a 'realization theory' that explains how the instantiation of the realizing facts guarantees the instantiation of the realized facts. What a realization theory provides, then, would be a way of seeing how the base necessitates what is based on it, and this would amount to an a priori derivation of the realized from a description of the realizer. But then this entails that the a priori entailment thesis has to hold, at least if we are committed to there being a realization relation between the physical and the phenomenal, which seems to be a core commitment of materialism.

Of course these two lines of reasoning are connected. One of the main reasons that materialists believe that the psycho-physical relation must be realization (or grounding), and not the weaker supervenience, is that supervenience alone would bring with it a brute necessity. If the phenomenal only supervened on the physical, but were not realized by it, then this would mean the physical would necessitate the phenomenal without there being an explanation of this necessity, and this is to be avoided if possible. So if materialism is committed to a realization relation holding between the physical and the phenomenal, and realizations entail explanatory connections, and these connections entail an a priori inference from the description of the base to a description of what is based on it, then it looks like the conceivability of zombies presents a real problem for Type B materialists after all.

Despite appearances, Type B materialists can maintain both that zombies are conceivable and that phenomenal consciousness is realized by, not just supervenient on, the physical facts. The key to reconciling these two claims lies in the dual nature of identity.

Where Chalmers locates semantics as the third vertex of the golden triangle that connects rationality to modality, I put identity instead.

By identity's 'dual nature' I mean that identity claims are normally a posteriori, established empirically, but, when true, have the stronger modal status of necessity. The necessity of identity, as both Kripke and Hume emphasized, is a priori: it is a matter of logic that everything is what it is and not something else. So if water is identical to H_2O, then it could not be anything other than H_2O. In other words, the schema, $\forall x \forall y(x=y \rightarrow \Box x=y)$, is a priori.

In fact, I would say that the entire phenomenon of a posteriori necessity comes down to this dual nature of identity. Rather than there being a complete break between epistemic and metaphysical modality, as some might interpret the results of the Kripke–Putnam thought experiments, in fact that connection remains secure. It is just that there is this one kind of necessary claim—an identity claim—which, due to its special nature, can be established by a posteriori reasoning. This dual nature of identity is in turn a straightforward consequence of the fact—made so much of by Frege—that it is possible for us to have two (or more) distinct representations for the same entity without realizing it. Once we pin a posteriori necessity on this dual nature of identity, we see how the Type B materialist can maintain both that phenomenal consciousness is realized by physical facts and also that zombies are conceivable.

Let us see how this works. The Type B materialist claims that the physical facts necessitate the conscious facts—that is, the latter supervene on the former. Yet, it is conceivable that all the physical facts are as they are and still there is no consciousness—zombie worlds are conceivable. How do we avoid being committed to brute necessity here? Well first, we identify phenomenal consciousness with some functional, or computational, or neural state (pick your favorite). Once that identification is made, we now have a redescription of phenomenal consciousness in (say) functional terms. When we compare our full description of the physical facts with our new description of phenomenal consciousness in functional terms we see (idealizing, of course) that the functional description follows a priori from the basic physical description. Hence we conclude that phenomenal consciousness is indeed realized by, and not merely supervenient on, the physical facts. Having established this realization relation, we see that the necessity in question here is not brute but quite explicable by an a priori derivation after all.

The idea here is that the necessity of phenomenal consciousness, given the physical facts, is indeed explicable, but just not under the description we usually employ to represent it. There are no brute necessities here, since there is a description for every phenomenon that is necessitated under which it is a priori derivable from some description of what necessitates it. It is the identity claims, themselves established empirically, but, as a matter of logic (and thus a priori), carrying modal import, that allow us to substitute the one description for the other and therefore explain the necessity in question; in this case, the supervenience of the phenomenal on the physical.

As I said above, it seems to me that all a posteriori necessities derive in this way from an empirically established necessary identity. So, take the proposition that water contains hydrogen. This is an a posteriori necessity that is not itself an identity claim. Still, one

can explain its necessity by appeal to an identity. First we discover that water is H_2O. Once we know they are the same thing, and therefore are so in every possible world, we are then rationally licensed to substitute the term 'H_2O' for the term 'water' in modal contexts. So from 'Necessarily water contains hydrogen' we get, by substitution, 'Necessarily H_2O contains hydrogen'. The latter statement is clearly a priori, and so not a brute necessity. Given the way we derived it from the former statement, this shows that the former statement is not a brute necessity either. The space of possible worlds is still governed by rational, a priori accessible principles. Brute necessity has been banished.

Of course the obvious response to this method of ridding Type B materialism of the curse of brute necessity is to object that all that has happened is that the bruteness of the necessity of the supervenience relation has been displaced onto the necessity of the identity relation. If identities are indeed the real source of a posteriori necessities, then they too require explanation. While the advocate of the conceivability argument has their favored explanation for these necessities, deriving them from the requisite conceptual analyses of the macro concepts, the (non-exceptionalist, neo-Russellian) Type B materialist leaves these identity claims brute and unexplained.

In reply, the Type B materialist can insist that nothing is really being left unexplained. Consider an identity like water is H_2O. There are two claims that might call out for explanation here. First, what explains its modal status, its necessity? Well, for this we have an a priori explanation, embodied in the a priori nature of the schema presented above. It is a priori that something is what it is and nothing else; it is just logic. Second, what explains its truth? Why is water H_2O? Well, this is something for which we do not expect an explanation. Call it 'brute' if you like, but it is a completely benign sort of bruteness, and not one that in any way undermines the constitutive connection between rationality and modality. After all, things are what they are; how can you explain that? You can explain how something came to exist (for concrete entities anyway), you can explain why it is rational to believe a certain identity claim, and you can explain how it is that this one thing happened to have two distinct representations of it, but it is not at all clear what it would mean to actually explain the identity itself. Things are what they are and nothing else—period.

Here is another way to put the argument between the conceivability challenger and the Type B materialist. If one wants to maintain some version of 'modal rationalism' (Chalmers's term), on which there is a constitutive connection between the a priori and the necessary, and thus brute necessities are banished, you have to come to terms with the (close to) universally acknowledged examples of a posteriori necessities somehow. This means that somewhere in the justification for these necessary truths an empirical element must be introduced. The question then is where to locate that empirical element so as not to undermine the constitutive connection between the a priori and the necessary.

What we have seen is that there are two ways to introduce this empirical element. The first way is the one advocated by the conceivability challengers to Type B materialism.

On this view, we have an a priori grasp of the satisfaction conditions for the relevant concept(s), and the empirical element comes into play in determining which entity in this world meets these conditions. So, for example, we know that water is whatever is the local entity that plays the watery role, and then it takes empirical discovery to determine that it happens to be H_2O. If we take this option, then, for all the reasons rehearsed by the advocates of the conceivability argument, there is no corresponding room for empirical discovery to enter into the justification for psycho-physical necessities, and so one cannot treat Type B materialism in the same way as other a posteriori necessities. Hence Type B materialism seems committed to brute necessities.

On the second way, the one presented above, there is no appeal to an a priori grasp of a concept's satisfaction conditions. What is a priori is just logic, which includes the logic of identity—specifically, that identities are necessary (if true), as in the schema presented earlier. Where the empirical element comes into play is in providing justification for the identity claim in question, for believing it true in the first place. So, for example, we do not identify water with H_2O by bringing to bear our a priori conception of the watery role and seeing which entity in the world satisfies it. Rather, we find that identifying water with H_2O is the best hypothesis to explain the behavior of water, and so is a straightforward application of inference to the best explanation. Once that identity is empirically established, we then get modal consequences from our a priori knowledge of the logic of identity. If viewed this way, then there is no obvious barrier to applying the same reasoning to the identification of phenomenal consciousness with some functional or physical state, and then we get the very same modal consequences as we did in the case of water. Again, the point is that empirical information has to come into play somewhere in the etiology of an a posteriori necessity. By locating the empirical component in the justification of the identity statement that licenses the redescription of the target phenomenon, whether it be redescribing water as 'H_2O' or phenomenal consciousness in physical/functional terms, the Type B materialist can employ the very same reasoning to avoid a commitment to brute necessities.

While I believe this is a perfectly plausible defense of Type B materialism in response to the challenge of the conceivability argument, it presents my overall position with a serious problem. At the outset I presented my commitment to two claims: First, I rejected the conceivability argument as an a priori refutation of Type B materialism. In Sections 18.2 and 18.3 I have defended that claim. But second, I also claimed that there is an explanatory gap between the physical and the phenomenal, and that one was warranted, by inferring to the best explanation, in rejecting materialism despite the defense of it just presented. However, given the argument just presented, that identities do not require explanations, how do I defend the claim that there is an explanatory gap after all? And if there is an explanatory gap, does that not mean I should accept the original conceivability argument? So what I need is an account of psycho-physical identities that leaves room for an explanatory gap but not in a way that provides support for an a priori refutation of materialism. I turn to that task in Section 18.4.

18.4 Explanation and Psycho-Physical Identity

The defense against the conceivability argument I have laid out for the Type B materialist crucially involves an appeal to the claim that identities do not require explanation.[6] This claim is crucial to the defense because psycho-physical identities are both necessary and, by hypothesis, not explicable, so if they did require an explanation they would constitute just the sort of brute necessities we want not to be committed to. But if identities do not require an explanation, where do we locate the explanatory gap?

Prima facie, the claim that identities are not apt for explanation makes good sense. Consider our standard example, water is identical to H_2O. Suppose someone asked, 'but why is water (identical to) H_2O'? I, for one, and not just because of my poor knowledge of chemistry, would not know how to answer such a question; it is unclear what the questioner is after. Why is water H_2O? Why is the Evening Star the Morning Star? Why is Mark Twain Samuel Clemens? In all of these cases it seems the only answer is, 'well, because they are'.

Of course the neo-Fregean can come up with an answer to these why questions. She can say that the answer to the question why water is H_2O is that H_2O satisfies the conditions conceptually constitutive of being water. That is what makes water H_2O. But as I see it, the question is an odd one and the fact that the (neo-Russellian) Type B materialist must consider such questions odd is not at all an embarrassment. In the end, of course, everyone will point to the fact that H_2O manifests the relevant superficial properties of water to justify the claim that they are identical. But for the Type B materialist pointing to these facts comes in by way of justifying acceptance of the identity claim, not explaining what makes it true.

This last point is worth emphasizing. Nobody denies that the reason we think water is in fact H_2O is that its being so explains its superficial behavior: why it is liquid at room temperature, freezes and boils at the temperatures it does, quenches thirst, etc. The disagreement concerns what role these facts play in establishing the identity claim. As I said above, for the neo-Fregean they play the role of determining which entity in the actual world (or the world of the scenario being considered) plays the role that is conceptually constitutive of being water. On this view it is a priori that water is whatever plays the watery role, so once one finds out which entity in the world in question plays that role, it just follows, a priori, that it is water. But for the neo-Russellian, the features that are collectively described as 'playing the watery role' are not conceptually constitutive of being water—nothing is, as on this view it is an atomic concept. However, the watery features are attributed to water as a matter of experience, and what we call 'the watery role' embodies extremely well-confirmed beliefs about the behavior of water. Thus, when we

[6] This section has been influenced by reading O'Conaill (unpublished) and subsequent correspondence.

find that H_2O explains these features, we have excellent, but still non-demonstrative reasons for identifying water with it.

So there is, then, a sense one can attach to the question 'why is water H_2O?' for the neo-Russellian, but it is not really a question about why water is H_2O, rather a question about why we should believe that water is H_2O. However this interpretation of the 'why' question will not support there being an explanatory gap when dealing with psycho-physical identities. To the question, why believe that (say) pain is c-fiber firing (or some functional state realized by c-fiber firing), the Type B materialist has an apparently easy answer: we should believe it because there is so much about the behavior of pain that is explained if we identify it with c-fiber firing. So on this picture it does not look as if there is room for the explanatory gap.

There is, however, another sort of question one can ask about an identity. Instead of why the identity holds—which, for the moment, we are granting is not an apt question—we can ask *how* the identity *could* hold. In other words, we can ask a 'how possible?', or 'how could it be?' question about an alleged identity.[7] So, to quote an example of E. J. Lowe (2000), suppose a Pythagorian tells you that standard middle-sized concrete objects, like tables and chairs, are in fact really numbers. One's query in response is not going to concern *why* that is the case, but rather *how it could be* the case. If an identity is claimed to hold and one cannot really see how it could, that certainly seems appropriately described as an explanatory gap. The idea is that explanations make their explananda intelligible, and what seems to be lacking with psycho-physical identities, according to this line of argument, is intelligibility. What we seek is a way to make the identity claim intelligible.

While I think in the end the explanatory problem with psycho-physical identity claims is of the 'how possible?' variety rather than of the 'why is it?' variety, the dialectic around this is complicated. So, consider the example above, the Pythagorian theory. If someone seriously proposed that, say, the chair I am sitting on is in fact identical to a number, I would of course reply that the very idea of the chair being a number is unintelligible. But what makes it unintelligible? Well, I might say the following. The chair has mass and is extended in three spatial dimensions, while numbers are abstract and lack such properties. In other words, when the claim that A is identical with B seems unintelligible, the lack of intelligibility can be attributed to one (or both) of A and B possessing properties that it is difficult to see how the other could instantiate. A number is not extended in space, has no mass, and the like, so how could it be identical to a chair that clearly instantiates these properties?

But now, let us try to apply this to psycho-physical identities. Someone proposes that pain is identical to the firing of c-fibers, or that the phenomenal experience of seeing red is identical to a pattern of firing in V4. One might then ask in response, 'but how could an experience of a color just be a pattern of neural firing?' Well, why could it not be?

[7] This is how O'Connail (unpublished) puts it. In early formulations of the explanatory gap I had also described the issue as making the identity 'intelligible', seeing 'how it could be true', which I take to be two ways of saying the same thing.

Presumably, one would appeal to certain properties of an experience that seem not to be attributable to a pattern of neural firing. For instance, experiences of red have a certain qualitative character that seems not appropriately attributed to neural states. If we know one side of the identity claim represents something that has a certain property and the other side represents something that does not have that property, then, by Leibniz's Law, we know the identity claim is false. So the explanatory question here is, how could a pattern of neural firing instantiate this reddish qualitative character? Explain that and then the how-possible question is answered (assuming this can be done for all the relevant properties).

The problem with this way of characterizing the explanatory gap is that the Type B materialist (e.g. Papineau 2002) will reply that we have misunderstood the original identity claim. It is not that there is some state that has a qualitative character and we are identifying that state with a neural state, which then commits one to making intelligible how the neural state could instantiate the qualitative character. Rather, the property in question—the qualitative character itself—is being identified with the relevant neural property. So the kind of how-possible question we want to ask, which relies on specifying a property instantiated by what is represented on one side of the identity that seems unintelligibly attributed to what is represented on the other side of the identity, cannot even be asked.

So here is the dilemma. If you ask a how-possible question about any alleged psychophysical identity, it is incumbent upon you to find some property clearly instantiated by one term that it seems unintelligible to attribute to the other term. But as soon as you present such a property, the reply will be that it is that very property itself that is being identified with its counterpart across the identity sign. So we seem to be left with only the why question, and that we already admitted, when it comes to identity claims, cannot be asked.

Yet, one wants to say, clearly there is a question here, and it is evident from the contrast we can draw between the psycho-physical case and the standard examples of theoretical identities. I really do not have any idea what someone has in mind if they ask either why water is H_2O or how possibly water could be H_2O.[8] However, when someone finds unintelligible the identity between a visual experience of color and some neural state I perfectly well understand what puzzles them.

In the past I have addressed this question in two ways. In Levine (2001) I distinguished between what I called 'gappy identities' and 'non-gappy identities'. A gappy identity is one for which the question 'why?' or 'how?' is intelligible, whereas a non-gappy one is one for which such questions did not really make sense. Of course this is really to name the problem rather than solve it. I am convinced there is a distinction here, but the distinction itself does not provide much insight into what it is. In Levine (2007) I leaned in

[8] Note that I would understand if someone thought, like Aristotle, that water was continuously divisible, and so therefore couldn't be swarms of H_2O molecules, which aren't. But then we would explain that water doesn't actually have that property; instead, it has the property of *appearing* continuously divisible. We can then explain how swarms of H_2O molecules could instantiate that appearance property.

Papineau's direction and acknowledged that what underlies the explanatory gap is really what he called an 'intuition of distinctness', but then, unlike him, argued that there were good reasons to take this intuition seriously and not consider it question-begging, as the standard explanation for the intuition, involving the phenomenal concepts strategy, did not work. Again, I think there is something right about this response too, but it still does not feel adequate to me. So let me try again.

Here is the structure of the problem. Consider property P, which supervenes on property B. If the necessitation of P by B is not to be brute and inexplicable, then there must be some representation of P, R,[9] such that one can derive P's instantiation from B's instantiation under those representations; that is, something like 'Bx → Rx' is a priori. But in order for this to be an explanation of how P, as originally represented (say, as 'P'), is necessitated by B, it must be that P = R is true.

Now, take 'P = R'. Suppose we cannot see how it could be true. Given our model of what it is to find an identity claim unintelligible, what this means is that P and R each have properties that it is hard to see how the other could instantiate. Let us say P has Q and it is difficult to see how R could have Q. But then maybe the problem is that Q is identical to S, R has S (or we can understand how it might have S), and so therefore it has Q. But then one might have a problem with the identity of Q and S, asking, again, how is that possible?

Now, on the model we have come up with for what it is to ask a 'how possible' question, questioning the intelligibility of the identity between Q and S would have to rest on specifying some further property, X, that Q has and for which it is difficult to see how it could be instantiated by S. It does seem that this has to give out at some point. Then the question is, can one just say that the identity does not make sense without appealing to yet further properties? If not, then why can the Type-B materialist not just keep proposing identities of this sort until we have no more properties to play the game with? At that point a brute 'intuition of distinctness' is all the anti-materialist can point to and it can easily seem question-begging.

In Levine (2007) I argued that given the 'core contrast' between standard theoretical identities and psycho-physical identities, it was incumbent on the materialist to explain why we have this intuition of distinctness, and not rest complacent with the claim that relying on a mere intuition is question-begging. I further argued that the phenomenal concept strategy does not explain this contrast, at least not so long as one is restricted to materialist implementations of phenomenal concepts, as of course Type-B materialists are.

A few paragraphs from the end of the paper, I said the following:

> One might say that there now is a second explanatory gap: between implementations of cognitive architecture and whatever it is about phenomenal concepts—in my terms, that they afford genuine cognitive presence to phenomenal properties—that is responsible for the original explanatory gap. If one thought the original

[9] We also need an appropriate representation of B, but let's just assume that 'B' (or what goes in for that schematic letter) is the one we want.

> explanatory gap was a problem and needed to be explained away, then one ought to be bothered by this one as well. . . Suppose I'm right that we can't now imagine how a materialist story of phenomenal concepts would go. No mere physical-causal mechanism can provide the kind of cognitive presence we seem to enjoy with respect to our phenomenal experience. So what is it we need? It seems to me that we need something like the old-fashioned relation of acquaintance. We are acquainted with our experience, and as acquaintance *presents* properties, not merely represents them, we find it difficult to integrate what is presented with what is only represented in a way that allows the latter to explain the former. If acquaintance itself cannot be explained in terms of physical-causal mechanisms, as I claim (at least so far) it can't, then we have to contemplate the possibility that it is a brute relation. If so, then the Materialist Constraint is violated, and materialism is false. (Levine 2007)

I think in the end the proper location for the explanatory gap is precisely at this second level, so it is not that there are two gaps, but rather the real source of the gap is the one identified in this passage. Here is how I want to put it now.

Let us assume we have reached the stage in questioning the intelligibility of an identity claim where there do not seem to be any further properties to appeal to. As in the model above, we hit the point where we have an 'intuition of distinctness' concerning Q and S. It is important to emphasize that the intuition of distinctness is not just an intuition to the effect that the properties in question are distinct, but that one does not see how they could be the same. After all, if it were not for the lack of intelligibility attaching to the identity claim, what would be the basis of the intuition?

To get concrete, what might be the relevant property to put in for 'Q' here? I think something like being a determinate *quality*, where reddishness, painfulness, itchiness are examples of what I mean by qualities. Neural firing patterns or causal roles, or whatever physico-functional properties are the proposed candidates for materialist identity claims, do not seem to have this feature of being qualitative. Now suppose the materialist finds some property of neural firing patterns, or functional roles, and says this is what being qualitative is and the response is, but how can that be? How can being qualitative be just identical to playing a certain functional role?

Given our initial assumption, there is now no other property of being qualitative that we can point to which does not seem intelligibly instantiated by the functional role (or neural firing pattern) property. We just hit bedrock and want to say that we just do not see how being qualitative could be that. The materialist wants to say at that point that if you cannot come up with another property to use in a Lebniz's Law argument against the proposed identity, then your resistance is merely intuitive and has no probative value. But I come back and say that indeed there does not seem to be another property to appeal to for a Leibniz's Law argument, and yet we cannot shake the feeling that there is something unintelligible about the identity. This requires explanation.

In any other case where we have a how-possible problem about an identity there is some further property to appeal to. Once those are exhausted, the intuitive problem disappears as well. So long as our concepts have run out of associations with other concepts and now function—à la the neo-Russellian view—as mere labels, there does not seem to

be any basis for intuitive resistance. So it must not be a matter of further properties, but rather the form of access to the property in question. Here is where the appeal to acquaintance comes in. It looks as if the kind of cognitive relation we bear to the contents of our experience is different in kind from that we bear to other sorts of properties, and it makes sense that it is due to this fundamental difference in modes of access that we persist in our intuitive resistance.

Now once we admit that the mode of access itself is a way of revealing difference, or creating an intuition of distinctness, we have to ask ourselves whether acquaintance is plausibly a physical (or physically realized) relation or not, and whether what is presented by that relation—that with which we are acquainted—is plausibly physical, or physically realized. At this point of course the materialist can still insist that they are identical to physical properties and we just cannot see it. I grant this is a consistent position and do not rule it out on a priori grounds. But I find it implausible, and think inference to the best explanation is that we have transcended the physical at this point.

I want to point out again, for emphasis, just where the move to acquaintance is doing the work I need to solve the problem I faced at the beginning of this section. My problem was that I could not provide an account of how one could find the proposed identity of qualitative character with a physico-functional property unintelligible—subject to a how-possible worry—unless there was yet another property of qualitative character that could support a Leibniz's Law objection. By appealing to a fundamentally different mode of cognitive access to qualitative character—acquaintance—I can make sense of the how-possible worry without recourse to another property. If conscious experience involves acquaintance with the contents of experience, and we are acquainted with nothing else, then this explains why the bare intuition of distinctness (that is, one that is not based on a Leibniz's Law objection) arises here and nowhere else. What explains this aspect of acquaintance? Well, this is, again, where I would use inference to the best explanation to support the claim that it is because, with acquaintance, we are not dealing with a purely physical phenomenon.

References

Alter, T. and Walter, S. (2007), *Phenomenal Concepts and Phenomenal Knowledge*. Oxford: Oxford University Press.

Biggs, S. and Wilson, J. (2017), 'The A Priority of Abduction', *Philosophical Studies*, 174/3: 735–58.

Block, N. and Stalnaker, R. (1999), 'Conceptual Analysis, Dualism, and the Explanatory Gap', *Philosophical Review*. 108: 1–46.

Chalmers, D. (2002), 'Consciousness and Its Place in Nature', in D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press.

Chalmers, D. (2006), 'The foundations of two-dimensional semantics' in M. Garcia-Carpintero and J. Macia (eds), *Two-dimensional Semantics: Foundations and Applications*. Oxford: Oxford University Press.

Chalmers, D. and Jackson, F. (2001), 'Conceptual Analysis and Reductive Explanation', *Philosophical Review*, 110: 315–61.

Gendler, T. and Hawthorne, J. (eds) (2002), *Conceivability and Possibility*. Oxford: Oxford University Press.

Kripke, S. (1980), *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Levine, J. (1998), 'Conceivability and the Metaphysics of Mind', *Noûs*, 32/4: 449–80.

Levine, J. (2001), *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press.

Levine, J. (2007), 'Phenomenal Concepts and the Materialist Constraint', in T. Alter and S. Walter (eds), *Phenomenal Concepts and Phenomenal Knowledge*. Oxford: Oxford University Press.

Levine, J. (2014), 'Modality, Semantics, and Consciousness', *Philosophical Studies*, symposium on David Chalmers, *The Character of Consciousness*, 167/3: 775–84.

Loar, B. (1997), 'Phenomenal States', in N. Block, O. Flanagan, and G. Güzeldere (eds), *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Lowe, E. J. (2000), *An Introduction to the Philosophy of Mind*. Cambridge: Cambridge University Press.

O'Conaill, D. (unpublished manuscript). 'Identity and the Explanatory Gap'.

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

Poland, J. (1994), *Physicalism: The Philosophical Foundations*. Oxford: Clarendon Press.

Putnam, H. (1975), 'The Meaning of Meaning', in K. Gunderson (ed.), *Language, Mind, and Knowledge*. Minneapolis: University of Minnesota Press, 131–93.