The Problem of Consciousness

David Papineau

[Forthcoming in *The Oxford Handbook of the Philosophy of Consciousness* ed Uriah Kriegel.]

…

The Explanatory Gap

There is a huge contemporary literature on physicalist views of the mind, covering a range of questions. How exactly should we define "physical"? Can mental properties be identified with basic physical properties, or should we instead embrace some version of "non-reductive physicalism", according to which mental properties supervene on, or are grounded in, or are otherwise constituted by basic physical properties, without being strictly identical to them? Do these non-reductionist options succeed in avoiding the epiphenomenalist threat that prompted physicalism in the first place? And so on.

However, we can by-pass all these issues here. This is because *any* version of physicalism about conscious states seems to generate pressing philosophical problems. Despite the strength of the argument for physicalism, the equation of the lived experience of perceptions, emotions, and pains with neuronal oscillations in the brain strikes many philosophers as effectively incomprehensible. As Thomas Nagel puts it in *The View from Nowhere* "We have at present no conception of how a single event or thing could have both physical or phenomenological aspects, or how if it did they could be related" (p 47). Or, in the more direct words of Colin McGinn, "How can technicolour phenomenology arise from soggy grey matter?" (1991, p 1.) To Nagel, McGinn, and many other philosophers, the idea that conscious states are at bottom physical seems obviously problematic.

This is the central problem of consciousness for contemporary philosophy. Arguments from causal closure provide compelling reasons to view conscious states as physical. But any such physicalist view of consciousness strikes many as little short of unintelligible. (Since this problem arises for all versions of physicalism, non-reductive as well as reductive, I shall often simply the exposition from now on by talking of conscious states as physical, or as identical to physical states; everything that follows will apply equally to physicalist positions that view conscious states as supervenient on, or grounded in, or constituted by, physical states.)

If we are to make any progress with this central problem of consciousness, we need to articulate the nature of the resistance to physicalism illustrated by the quotations from Nagel and McGinn. One useful way to do this is to compare putative mind-brain identities with similar scientific identity claims in other areas. When we are told that common salt is NaCl, or lighting is atmospheric electrical discharge, we happily accept these claims as telling us about the underlying physical nature of these everyday phenomena. But when we are told that pains are the firing of prefrontal nociceptive-specific neurons, or that visual experiences of red are neuronal oscillations in the V4 area of the visual cortex, we react quite differently. Even after we are given this information, we still want to know *why* those brain states are accompanied by those feelings. Why do the nociceptive-specific neurons, or the oscillations in V4, feel like that, rather than some other way, or no way at all? As Joseph Levine has put it, mind-brain identities seem to leave us with an *explanatory gap*, in a way that other scientific identities do not (Levine 1983). We remain puzzled about why the brain states give rise to the feelings, in a way that we don't feel puzzled about why NaCl gives rise to salt, or electrical discharges to lightning.

Now, as a social or psychological phenomenon, the existence of an explanatory gap is quite uncontentious. There is no doubt that most people react to mind-brain identity claims with demands for further explanation, in a way they don't react to other scientific identity claims. However, the philosophical significance of this social fact is far less straightforward. Philosophers disagree widely about the source of the reaction and about what, if anything, it implies about the relation between conscious and physical states.

There are two distinct questions here. The first is a psycho-social question. What is the *source* of the explanatory asymmetry? Why do people feel that mind-brain identities, unlike scientific identities, leave something unexplained? The second is a philosophical question. What *follows* from this explanatory asymmetry? Does the puzzlement occasioned by mind-brain identities imply that there is some deficiency in the physicalist view of consciousness?

I shall address these issues in turn. The next three sections will be devoted to the source of the explanatory asymmetry. After that I shall turn to the philosophical implications.

The Derivability Gap

The psycho-social question first. My own view on this is a straightforward one. I think that the feeling of an explanatory gap is simply an upshot of the fact that we all—including professed physicalists like myself—find mind-brain identities almost impossible to believe. Even after we are shown plenty of evidence that pains and nociceptive-specific neuronal firing always accompany each other, and have the same causes and effects, we still intuitively resist the conclusion that they are identical. How could that urgent feeling possibly be one and the same as neuronal activity, we ask? We find it hard to escape the

spontaneous dualist thought that the feeling and the physical state are not one thing, but two different states that somehow invariably accompany each other.

And, to the extent that we do think this, then of course we feel a need for more explanation. Why is the neuronal activity accompanied by the nasty feeling of pain, rather than a pleasant sense of floating, say? Indeed, why is it accompanied by any feeling at all? Once we slip into the dualist way of seeing things, we cannot avoid a range of demands for further explanation. (See Papineau 2010.)

Most philosophers currently working on consciousness, however, take a quite different and less straightforward view of the feeling of an explanatory gap. In their view, this feeling is not a consequence of an intuitive resistance to physicalism. Rather, it stems from an internal feature of the way we think of conscious states, and would persist even if we were able fully to embrace physicalism.

This mainstream view attributes the feeling of an explanatory gap to the impossibility of *deriving* mind-brain identities a priori from the physical facts. This is supposed to mark a contrast with the scientific cases. While we can often derive scientific identities a priori from the physical facts, so the thought goes, we can't so derive mind-brain identities, and this creates a feeling of puzzlement about them.

The reason for the difference, on this mainstream account, lies in the different ways in which we ordinarily *conceive* of scientific properties and conscious properties. Consider our everyday concept of *common salt*. According to the mainstream view, we think of salt as the stuff that is white, crystalline, granular, with a distinctive taste, that dissolves in water, and is found in the oceans. Now imagine someone who has a fully detailed account

of the physical make-up of the world, in terms of the distribution of matter, arrangement of elementary particles, the deployment of fields, and so on. In principle, such a person could arguably put this knowledge together with their prior conceptual grasp of *salt* to figure out that salt must be NaCl, on the grounds that NaCl is the stuff that fits the conceptual requirements for salt—white, crystalline, . . .

However, we can't do this with pain, say, or with visual experiences of red. The problem is that our everyday concepts of *pain* or *visual experience* don't pick out their objects via some descriptive role, like white, crystalline, . . . but in terms, so to speak, of what the states feel like. In the first instance, we think of conscious states directly, by focusing on the feelings involved, and not as the states, whatever they may be, that play some descriptively specified role. And this blocks any a priori derivation of mind-brain identities from the physical facts, of the kind that is arguably available for identities like salt = NaCl. Scrutinize the physical facts as much as you like, and they won't tell you that pains are the firing of prefrontal nociceptive-specific neurons. Since we don't think of pains in terms of some specified role, but in terms of the feelings involved, there is no way to connect the physical facts with the phenomenon of pain.

Given this, our knowledge of mind-brain identities can only be based on some kind of a posteriori abductive inference, rather than a principled a priori demonstration that a certain physical state fills some specified role. For example, we might observe that pains occur whenever prefrontal nociceptive-specific neurons fire, and vice versa; we might also note that, if pains were the firing of nociceptive-specific neurons, then this would account for a number of other observed facts about pain, such as that it can be caused by trapped nerves, and can be blocked by aspirin; and we might conclude on this basis that pains are indeed identical to the firing of nociceptive-specific neurons.

Still, to repeat, there is no question of deriving this identity a priori from the physical facts, by showing that the nociceptive-specific neuronal firing fills the pain role – for we don't think of conscious pains in terms of roles to start with. And this, says the view under examination, is why we feel an explanatory gap in the mind-brain case. The peculiarly direct nature of our concepts of conscious states stops us deriving mind-brain identities a priori from the physical facts.

Doubts about Derivability

This lack-of-derivability account of the source of explanatory-gap feelings is widely taken for granted in contemporary philosophy of mind. Despite this, I think it is clearly mistaken, and shall explain why in a moment. One part of the story, however, is relatively uncontentious. This is the idea that we have direct, non-descriptive concepts of conscious states that preclude any a priori derivation of mind-brain identities from physics.

Some initial mid-twentieth-century versions of physicalism did not accept this, and so held that mind-brain identities could indeed be read off from the physical facts. But this stance was dealt a critical blow by Frank Jackson's "Knowledge Argument" (Jackson 1986). Jackson pointed out that someone who has never experienced colours could be in possession of all the physical facts about colour vision, and yet not "know what it is like" to see something red. The overall philosophical significance of Jackson's argument is a complex matter, to which we shall return in due course. But it is pretty much agreed on all sides that, at a mimimum, Jackson's argument does demonstrate the existence of a special range of "phenomenal concepts", ways of thinking about conscious states directly, in terms of the feelings involved, which are normally only available to subject who have experienced those states

themselves, and which block any physics-based derivation of mind-brain identities.[1]

So far so good. We can't derive mind-brain identities a priori from the physical facts. Still, is this really the source of the feeling that the identities leave something unexplained, as claimed by the suggestion currently under examination? This suggestion faces an obvious objection. Plenty of other identities similarly can't be derived from physics, but generate no corresponding impression of an explanatory gap.

After all, as the above remarks make clear, a priori derivations from physics will be blocked whenever we have concepts that refer directly, rather than by association with some described role. On the face of things, phenomenal concepts are by no means the only such cases. Proper names ("Cary Grant"), demonstrative constructions ("that dog"), and simple terms for observable properties of objects ("round") are all arguably terms that refer directly, rather than by description. Given this, when we accept identity claims involving these terms (such as "Cary Grant = Archie Leach", or "that dog = her pet", or "round = locus of constant distance from some point"), it can only be on the basis of an abductive inference from direct empirical evidence, such as that the two things in question are found in the same places and the same times, and are observed to bear the same relations to other things, not because we can deduce the identities a priori from the physical facts.

[1] Among the exceptions who resist phenomenal concepts are David Lewis (1988) and Daniel Dennett (1992). Curiously, they have now been joined by Jackson himself, who has come to view dualism as untenable while continuing to maintain that it would follow from the existence of phenomenal concepts (2007). Other recent writers reject phenomenal concepts on the grounds that no concepts can constitutively depend on prior conscious experiences (Ball 2009, Tye 2009); in my view, this argument set the standards for phenomenal concepts too high.

Yet we feel no explanatory disquiet when presented with these identities. Even though they must perforce be based on some form of abductive inference, for lack of any descriptive roles associated with the relevant terms, they certainly don't leave us with a feeling that something has been left unexplained.

Come to think of it, it is doubtful that many scientific identities are based on anything more than abductive inferences either. When nineteenth-century scientists first figured out that salt is NaCl, they certainly didn't do so by inferring a priori from basic physical theory that NaCl molecules will appear white, form crystals, dissolve in water, and so on, and hence concluding that NaCl must be the substance that fits the specifications for salt. The sub-atomic understanding required for such derivations was more than a century in the future. Rather the original scientists simply noted that NaCl molecules were present whenever salt was, and vice versa, and had some of the same causes and effects, and identified them on that basis.[2]

The same goes for the identification of lightning with electrical discharge, or consumption with tuberculosis infection, or nearly all other scientific identities. Scientists didn't derive these identities a priori from physical theory, but based them largely on simple observations of co-occurrence and matching casual relations to other things.

Yet this didn't make the scientists feel something was left unexplained. Even though the identities were based on abductive inference, rather than derived a priori from the physical facts, the scientists weren't left unhappily puzzled about why NaCl gives rise to salt, rather than to something else, or to nothing at all, or why

[2] In any case, a derivation from the strictly *physical* facts alone was never really on the cards, given the presence of such observational terms as *white* in the conceptual role of *salt*. See Levine 2010.

lightning arises from electrical discharge, or consumption from tuberculosis infection. Once more, the absence of a priori derivations didn't seem to engender any feelings of things left unexplained.

Perhaps these counterexamples are less than conclusive. Still, there are further grounds, apart from possible counterexamples, for doubting the mainstream thesis that the feeling of an explanatory gap arises from our inability to derive the relevant identities a priori from the physical facts.

Note how this mainstream account implies that *something is left unexplained* when we embrace mind-identities on the basis of abductive inferences, something that *does get explained* when we supposedly derive scientific identities from the physical facts. But what exactly is that? What exactly does get explained, according to the mainstream account, when we derive scientific identities a priori from the physical facts, but not when we embrace mind-brain identities on the basis of abductive inferences?

One first thought might be that it is the identities themselves that are explained. We can explain *why* salt is NaCl, or *why* lightning is electrical discharge, once we can derive the identities a priori from the basic physical facts, in a way these identities would be left unexplained if we simply based them on abductive inferences.

But this seems odd. We don't normally regard identities as in need of explanation. Since they are necessary, they could not have been otherwise, and didn't need anything to make them so. (To repeat a familiar example, when we discover that Mark Twain = Samuel Clemens, we might reasonably ask *why* he had two names, or *why* nobody told us before. But it would make no sense to ask—*why* was Mark Twain the same man as Samuel Clemens? Block 1978.)

A second thought would be that it is not the identities per se that get explained, but the behaviour displayed by everyday kinds. So, for example, once we can derive the identity of salt with NaCl a priori from the physical facts, then we will in principle be able to explain why salt displays such defining characteristics as whiteness, crystallinity, solubility in water, and so on, in a way we can't if the identity is based on brute correlations.

But this second thought does not hold water either. Sure, we can explain the behavior of salt if we derive its identity with NaCl a priori from the physical facts, by appealing to our physical understanding of how NaCl molecules work. But, by just the same coin, we can explain the behaviour of pain if we accept its identity with nociceptive neuronal firings on the basis of an a posteriori abductive inference. As I observed above, an identification of pain with the firing of prefrontal nociceptive-specific neurons, even if based on an abductive inference, will happily allow us to explain such things as why pain is caused by trapped nerves, or why it is relieved by aspirin.

So, once more, it doesn't look like the impression of an explanatory gap can really be due to a lack of a priori derivation. Such derivations don't allow us to explain anything that can't be explained without them.

Much of the contemporary literature on the "explanatory gap" simply reads this phrase as referring to the impossibility of deriving mind-brain identities a priori from the physical facts. But we have now seen that, in truth, this understanding quite fails to answer the psycho-social question of why mind-brain identities leave most people with the feeling of an *explanatory* gap. For a start, people don't seem to have any feeling of non-explanation with other identities that cannot be derived a priori from the physical facts. Moreover, nothing extra would seem to be

explained when we can derive identities a priori from the physical facts.

In the end, though, there is an even more powerful reason for rejecting the idea that the feeling of an explanatory gap is something to do with a priori underivabilty. This is the availability of the alternative account mentioned earlier, an account that avoids all the difficulties raised in this section.

The Intuition of Distinctness

On the alternative account I favour, the issue is not that we feel that something still remains to be explained *after* we have accepted mind-brain identities. It is rather that we all find mind-brain identities very difficult to accept in the first place.

As I observed above, even after we are given all the abductive evidence, we still find mind-brain identity claims almost impossible to believe. We cannot resist the dualist conviction that conscious feelings and the physical brain states are two different things. And this, in my view, is the real reason why we feel a need for further explanation. We want to know why the neuronal activity is accompanied by *that* conscious feeling, rather than by some other, or by no feeling at all. Our dualist intuitions automatically generate a hankering for further explanation.

On my diagnosis, then, the demand for explanation arises, not because something is lacking in physicalism, but because something is lacking in us. Even after we are shown the arguments for physicalism, and are perhaps moved to embrace physicalism at a theoretical level, we continue to experience the pull of the dualistic perspective, and so intuitively feel that something remains to be explained.

If only we could fully embrace physicalism, this diagnosis suggests, the feeling of an explanatory gap would disappear. If we could full accept that pains are nociceptive-specific neuronal firing, then we would stop asking why “they” go together—after all, nothing can possibly come apart from itself. The feeling of a gap is simply a corollary of the intuitive grip of dualism.

From this perspective, then, a properly thorough-going physicalism promises to dissolve “the problem of consciousness”. The committed physicalist will simply deny that any puzzle is raised by the fact that it feels painful to be a human with active nociceptive-neurons. What shouldn’t it feel like that? That’s how it turns out. Why regard this as puzzling?

Note how my diagnosis in terms of intuitive dualism offers a far better account of the feeling of an *explanatory* gap than the appeal to lack of a priori derivability. For a start, it is now clear why we feel something has been left *unexplained*—we want to know specifically why brain states give rise to extra conscious states. Moreover, the feeling of a gap is now specifically about mind-brain relations, and so there’s no puzzle about why we don't feel it in other cases where a priori derivability is blocked.

By way of further support for the idea that the feeling of an explanatory gap stems from intuitive dualism, we need only attend to the phraseology normally used to discuss the relation between mind and brain. Brain processes are said to “generate”, or “yield”, or “cause”, or “give rise to” conscious states. (“How can technicolour phenomenology *arise from* soggy grey matter?”) These expressions are common currency in writings on consciousness, including by thinkers who say they are no dualists. But the phraseology itself is not consistent with physicalism. Fire “generates”, “causes”, “yields” or “gives rise to” smoke. But NaCl doesn’t “generate”, “cause”, “yield” or “give rise to” salt. It *is* salt. The point is clear. To speak of brain processes as “generating”

conscious states, and so on, only makes sense if you are implicitly thinking of the conscious states as separate from the brain states.

If further evidence is needed, consider our intuitive reaction to whether zombies are possible. Could a being share all your physical properties but have no conscious life? Everybody's first thought is, "Sure. Just duplicate the physical stuff and leave out the feelings."

Reflective physicalists will of course realize, on second thought, that they must deny that this is really possible. (If conscious states *are* physical states, the "two" cannot come apart.) But it is the initial reaction that I want to focus on here. Compare our response to the idea of Marilyn Monroe existing without Norma Jean Baker, say. I take it that our initial reaction to this suggested possibility would be puzzlement. What are we being asked to posit? That she exist without herself? That makes no sense.

This contrast is a reflection of our intuitive dualism. Zombies strike us as initially possible simply because all of us, physicalists included, intuitively think of conscious feelings and physical states as distinct existents. If we fully embraced the idea that they are one and the same, then we would find the idea of zombies simply puzzling. How could there be nociceptive-specific neuronal firing without pains? What are we being asked to posit? That the state exist without itself? That wouldn't make any more sense than Marilyn Monroe without Norma Jean Baker. (Cf Papineau 2007.)

If the feeling of an explanatory gap stems from our intuitive dualism, as I have been arguing, then the obvious next question is about the cause of these persistent dualist thoughts. Why do dualist ideas maintain such a firm grip, even on thinkers who are fully persuaded of the strength of the arguments for physicalism?

Plenty of possible answers to this question offer themselves, but before considering them I would like first to return to the issue left hanging earlier, namely whether the feeling of an explanatory gap is associated with any good *arguments* against a physicalist view of consciousness. After all, one possible explanation for why many people feel intuitively convinced that physicalism is false might be that they can all see that there is a strong argument against it.

Of course, even if there were a good philosophical argument against physicalism, it might not be the reason most people instinctively reject physicalism; the argument might not be apparent to them. But, even so, it will be useful to get clear about the nature of the arguments against physicalism, before discussing the possible causes of persistent dualist intuitions.

Arguments Against Physicalism

The best place to begin assessing the argumentative case against physicalism is with Jackson's "Knowledge Argument".

As I explained earlier, Jackson's argument hinges on the observation that someone could know all the physical facts about colour vision, and yet not "know what it is like" to see something red. And, as I said, this observation is generally agreed to demonstrate the existence of a special range of "phenomenal concepts" that refer directly to conscious states and are normally only available to subject who have experienced those states themselves.

Jackson original intention, however, was not just to argue for an extra set of phenomenal *concepts*, but in addition for an extra set of phenomenal *properties*. That is, he was arguing for the dualist

conclusion that conscious states are metaphysically distinct from and additional to any physical states.

Even so, once phenomenal concepts are on the table, physicalists would seem to have a ready initial response to his argument. They can say that "not knowing what some conscious states are like", even when you are completely knowledgeable about the physical facts, is simply a matter of not being able to represent certain physical states (the relevant conscious states) *using phenomenal concepts*. Once you know all the physical facts, then you know about all of reality. If, despite this, you still "don't know what some states are like", that's just a matter of your not being able to represent those states in the special direct way that only becomes available once you are possession of the relevant phenomenal concepts.

From the point of view of physicalists who take this line – "a posteriori" physicalists – we thus have two distinct kinds of concepts that refer to conscious states. On the one hand are phenomenal concepts – like *pain* or *seeing something red* – that pick out their referents directly, in terms of what they feel like, so to speak. And on the other are physical concepts – like *nociceptive-specific neuronal firing* or *oscillations in V4* – that refer to just the same states in terms of their physical nature. Scientific investigation can then show us that the former concepts pick out the same things as the latter ones, just as it establishes such other a posteriori identities as *salt = NaCl*, or *lightning = atmospheric electrical discharge*.

However, a second line of anti-physicalist argument now comes into play. This focuses on the particular nature of phenomenal concepts, and contends that certain features of these concepts are incompatible with their referring to physical states.

The basic thought is that, if physicalism were true, the *directness* of phenomenal concepts ought to render its falsity inconceivable – yet it doesn't. Consider once more a "zombie", a being who shares all my physical properties yet has no conscious life. Physicalists must deny that zombies are *possible*, given that the mind is ontologically inseparable from the brain. But a posteriori physicalists have no choice but to allow that they are at least *conceivable*. (If phenomenal concepts refer directly, and my feelings are therefore not a priori derivable from my physical properties, then there's no conceptual contradiction in ascribing a being all my physical properties, but denying it my conscious ones.)

The argument against physicalism now hinges on the thesis that impossibilities are only conceivable when presented using concepts that refer *indirectly*. For example, take *salt = NaCl*. Even though this couldn't be otherwise – salt *is* NaCl – someone can certainly conceive of (indeed believe in) NaCl not being salt. However, according to the argument at issue, they can only do this because they are thinking of salt at second hand, as *the* substance, whatever it is, that is white, crystalline, . . . This way of thinking leaves it open whether or not salt is in fact NaCl, and thus whether it is necessarily identical to that substance.

But a phenomenal concept like *pain* isn't indirect in this way. Phenomenal concepts don't pick out their referents indirectly, by some association with a role, as with *salt*, but directly, in terms of what they are like. So there is no room, the argument goes, for claims made using phenomenal concepts, such as *nociceptive-specific firing is pain*, to be necessarily true, yet appear conceivably false. If this claim were true, it would have to be a priori. Yet it isn't.

The crucial premise in this argument is that necessary facts can only appear conceivably false when they are presented using

indirect concepts. Or, putting it the other way around, once everything is formulated directly, no necessary truths will appear conceivably false. In short, the crucial premise is that direct concepts are *revelatory*, in the sense of displaying all the necessary properties of their referents a priori. (And then the anti-physicalist reasoning goes: the concept *pain* is direct, so it must be revelatory; so, if my physical nature necessitated my pains, this ought to be knowable a priori; but it's not; therefore pains can't be physical.)

A posteriori physicalists deny the crucial premise of this argument. They don't accept that direct concepts are always revelatory. Directness is a semantic matter—the concept picks out its reference directly, rather than as the item that satisfies some descriptive role. Revelatoriness is epistemological—the concept renders all necessary features of its referent a priori knowable. A posteriori physicalists insist that the former doesn't imply the latter.

In particular, they hold that phenomenal concepts are direct but not revelatory. They accept that phenomenal concepts are direct. And as physicalists they of course hold that pains have a physical nature. But they deny that this essential feature of pains must be revealed to us by the phenomenal concept *pain*. You can grasp this direct concept fully, yet not appreciate that pains are necessitated by the relevant brain processes.

An extensive literature is devoted to the question of whether all direct concepts are revelatory, and all directly formulated necessary claims are therefore knowable a priori. (See e.g. Block and Stalnaker 1999, Chalmers and Jackson 2001, Chalmers 2002, Levine 2001 2010.) A posteriori physicalists and other opponents of this thesis contend that there are plenty of counterexamples. What about identity claims involving proper names, indexical constructions, or observational concepts – "Cary Grant = Archie Leach", "that dog = her pet", "round = locus of constant distance from some point"? Earlier I cited the apparent a priori underivability of such claims from the physical facts as an argument against attributing the explanatory gap to this kind of underivability. In the present context, the same cases offer putative examples of directly formulated necessities that aren't a priori knowable.

At this point the arguments get messy. Anti-physicalists respond that, despite the prima facie absence of descriptive content, the terms in question should properly be understood as functioning indirectly, and that this this is why they do not reveal the identity claims involving them a priori. Some physicalists counter by questioning the way their opponents are drawing the distinction between direct and indirect terms. An alternative physicalist strategy is to grant that in general direct concepts are revelatory and directly formulated necessities a priori, and that a posteriori phenomenal mind-brain claims are therefore an exception to this rule, but maintain that there is nothing wrong with that. And so on. (See Levine 2001 ch 2.4.)

Fortunately, it is possible to cut through much of this dialectic. What really matters for the anti-physicalist argument is whether *phenomenal concepts* are revelatory, not any more general thesis about some wider category of "direct" concepts. The anti-physicalists say that phenomenal concepts are revelatory, and in particular that they reveal conscious states not to be physical. Physicalists respond that there is no reason to suppose that phenomenal concepts have the power to reveal such things.

Given this, it makes sense for us to address the revelatoriness of phenomenal concepts head on, and by-pass the further issue of whether this can be seen as a special case of some more general principle involving direct concepts. As far as the anti-physicalist

argument goes, all that matters is the workings of phenomenal concepts themselves. (Cf Nida-Rümelin 2007, Goff 2011.)

At an intuitive level, it is certainly not implausible that phenomenal concepts are revelatory. Consider what it's like to think at first hand about a stabbing pain, or a visual experience of seeing something red. Does not such thinking acquaint you with the very nature of these conscious states? It certainly seems as if such phenomenal thinking lays bare all essential aspects of the relevant experiences.

A posteriori physicalists will respond that appearances are deceptive. We should not be distracted, they will say, by the close association between phenomenal thinking and the experiences being thought about. Often the experience itself (the pain, the awareness of red) is present when we think about it phenomenally. In other cases, an imagined version of the experience (a "faint' copy", as Hume put it) accompanies our phenomenal thinking. And, because of this, it can seem that everything is revealed. A version of the experience is right there, before our minds. How can anything essential remain hidden?

But it is one thing, physicalists will object, to *have* an experience. It is another to *know* everything about its nature. Phenomenal thinking might characteristically *give* us the experience, in the sense that we undergo some version of it while thinking about it. But this doesn't mean it tells us everything about its nature. In particular, it doesn't mean it will reveal that the experiences are at bottom physical, if they are.

Moreover, the physicalist can continue, there is something deeply mysterious about the idea that merely thinking about something can reveal all its necessary properties. Of course, in the case of complex concepts with internal structure, mere thinking can deliver analytic knowledge; for example, someone who possesses

complex concept *square* can work out, just by analysing this concept, that *squares have four sides*. But this model does not seem relevant to the putative power of phenomenal thinking. Phenomenal concepts like *pain* or *seeing something red* do not seem complex; nor, correspondingly, do anti-physicalists maintain that the non-physicality of their referents is an analytic consequence of their internal structure.

Perhaps anti-physicalists can appeal to a different model. Instead of invoking analytic knowledge, they can construe phenomenal thinking as a kind of direct acquaintance, appealing to the point that such thinking is characteristically accompanied by versions of the experiences thought of. The idea would be that we find out about phenomenal states by thinking about them introspectively. We scrutinize our experiences internally, and thereby uncover their nature.

But the mystery remains. Introspection is certainly able to tell us what experiences we are having, and various other things about them. But why should it be guaranteed to tell us about *all* their necessary properties? How is that supposed to work? Any normal information-delivering process is inevitably fallible and only partially informative about the nature of its objects. To hold that introspection is guaranteed to reveal all necessary properties of experience would seem to take us beyond the realm of naturally explicable faculties.

Neutral Monism

Suppose for the moment that the argument from revelation did hold water. This would scarcely leave the anti-physicalist in a comfortable position. As I observed earlier, modern scientific findings seem to leave epiphenomenalism as the only viable alternative to physicalism. Yet the epiphenomenalist relegation of

conscious states to inefficacious causal "danglers" is not an attractive option. If this is where the argument from revelation ends up, that would itself be a reason for thinking it must have gone wrong somewhere.

But perhaps there is another way out. An increasing number of contemporary philosophers favour an alternative view, known as "Russellian monism", which offers a way of embracing the argument from revelation while avoiding the entanglements of epiphenomenalism. In effect, this position aims to maintain the causal significance of phenomenal states by viewing both the phenomenal and the physical as grounded in some more fundamental reality.

Let us go back to the argument from revelation. This said that a truth of the form *pains = nociceptive-specific firing* can only be conceivably false if it is formulated in indirect terms. The route from this to Russellian monism hinges on the thought that perhaps it is *nociceptive-specific firing* that is the indirect term, rather than *pain*.

So far I have not queried the idea that physical terms like *nociceptive-specific firing/NaCl/electric discharge* are direct and revelatory. But there is no reason to take this for granted. A standard account of scientific terms has them referring via theoretical descriptions – to that property, or quantity, that plays such-and-such a theoretically specified role. (So for example, *mass* might be equated with *that quantity that is inversely proportional to acceleration and obeys the law of gravitation*.)

This now offers a different way of squaring the conceivable falsity of *pains = nociceptive-specific firing* with the principle that necessary truths can only be conceivably false if formulated in indirect terms. Suppose that the term *nociceptive-specific firing* refers indirectly to that underlying property, whatever it is, that

plays the role specified by neurophysiological theory. Russellian monism now view the conscious feeling of pain as itself grounded in this underlying property. (Russell 1927.)

This allows us to account for the conceivability of zombies, beings who have nociceptive-specific firing but no pains, as possible beings in whom the relevant theoretical role is filled, not by the underlying property that constitutes pain in the actual world, but by some different and non-conscious property. Since we are thinking of the nociceptive-specific firing only indirectly, as the filler of a theoretically specified role, this leaves it open that this role could possibly be played by something other than its actual filler, indeed by something that fails to constitute any conscious feeling at all.

At the same time, this Russellian move promises to eliminate any worries about the epiphenomenality of pain. After all, pain is now constituted by a basic property, the property that fills the *nociceptive-specific firing* role in the actual world. At first pass, such basic properties look like just the kind of items to enter into fundamental causal relations.

This Russellian position is often associated with some version of the *panpsychist* doctrine that consciousness permeates all parts of the natural world. For some thinkers, this further commitment is motivated by the thought that our introspective awareness of our conscious experience is the only point at which we are directly acquainted with the underlying nature of reality. Since introspection shows reality to be conscious in all cases where its underlying nature is revealed, the thought continues, we should therefore conclude that it is conscious throughout. (Goff 2017.)

A further motivation for panpsychism derives from a perceived need to *explain* the consciousness that is present in beings with brains like ours. Russellian monists accept the orthodox view that

the underlying physical processes that constitute our conscious life are complex, and in particular that they are built up from the same simple components (fundamental field and particles) that compose the rest of nature. Given this, many feel that it would be mysterious for consciousness to emerge in complex brain processes if it were not already present in the simple parts. (Cf Strawson 2003.)

Despite Russellian monism's current popularity, it is questionable whether it marks any real advance on ordinary a posteriori physicalism. On further analysis, it turns out to leave us with many of the same issues, and moreover to generate a number of problems of its own

An initial difficulty relates to the explanation of macroscopic conscious states in terms of their microscopic parts. Even if the microscopic components are credited with some conscious nature, this will presumably be different in kind from the conscious nature of the wholes they compose. So why is the relation between the conscious parts and the differently conscious wholes any less mysterious than the supposedly puzzling emergence of conscious wholes from non-conscious parts? (Stoljar 2006.)

A converse puzzle involves our phenomenal knowledge of macroscopic conscious states with microscopic parts. If phenomenal concepts reveal all the necessary properties of their referents, then why do they not show pains and other conscious states to be composite? If some state is built from parts, then this is presumably part of its nature. Yet introspection presents conscious states like pains as simple and unified, not composite. (Lockwood 1993.)

A further worry is that Russellian monism seems to end up flirting with the very epiphenomenalism it is designed to avoid. It is

essential to the Russellian position that the *nociceptive-specific firing* role, say, might possibly be filled by a number of different underlying states, including ones that have no conscious nature (as in the zombie version of me). But now it looks as if the conscious differences between these alternative fillers make no difference to their causal powers. After all, by hypothesis these different fillers all display just the same behaviour and conform to just the same scientific laws. If, in addition, the fillers involve variations in consciousness, these variations would thus seem condemned to causal inertness.[3]

Finally, and relatedly, the general metaphysical position on which Russellian monism rests is itself highly contentious. As the Russellians see it, scientific terms are non-revelatory because the specification of a theoretical role leaves it open which underlying entity fills that role. But it is not obvious, to say the least, that we should accept this thesis. Consider the case of mass. As I said, science arguably picks this out as *that quantity that is inversely proportional to acceleration and obeys the law of gravitation*. From the Russellian perspective, then, there is another possible world, just like the actual world, save that some different quantity, *schmass*, plays the mass role there. But this seems a perverse commitment. Surely that would simply be another world that contains *mass*, the same quantity as is present in our world.

This is not the place to resolve the debate about the metaphysical relation between properties and laws. (Cf Bird 2007.) Still, on the face of things, the more natural view would seem to be that basic scientific properties are necessarily attached to their nomological roles. Fix the profile of laws that governs the entity, and you have

[3] For this line of objection see Howell 2015, and for a Russellian response see Alter and Coleman 2018.

fixed the entity itself[4]. Why multiply complexity unnecessarily by positing differences that have no further consequences?

All in all, then, Russellian monism seems to generate more problems than it solves. In my view, we would do better to stick with simple a posteriori physicalism, and forget about the supposed argument from relevation. Abductive evidence establishes certain phenomenal-physical identities. Even if both the phenomenal and physical concepts involved pick out their referents directly, the conceivably falsity of these identities does not discredit them. Why ever should we suppose that directly referring terms will reveal all the necessary features of their referents a priori?

…

References

Alter, T. and Coleman, C. 2018 “Panpsychism and Russellian Monism” in Seager, W. ed *Routledge Handbook of Panpsychism* London: Routledge

Armstrong, D. 1968 *A Materialist Theory of the Mind* London: Routledge and Kegan Paul

Ball, D. 2009 “There Are No Phenomenal Concepts” *Mind* 118:935-962

Alexander Bird 2007 *Nature's Metaphysics. Laws and Properties* Oxford: Clarendon Press, Oxford.

Block, N. 1978 "Reductionism: Philosophical Analysis" in *Encyclopedia of Bioethics,* London: Macmillan

Block, N. 1995 "On a Confusion about a Function of Consciousness" *Behavioral and Brain Sciences* 18: 227-87

[4] This is not to deny that some coarse-grained theoretical roles—that of an electrical insulator, say—can be variably realized by different states of affairs with different fine-grained specifications. But that isn’t enough for the Russellian monist, who needs even the most fine-grained theoretical roles to be variably realized.

Block, N. 2007 “Consciousness, Accessibility, and the Mesh Between Psychology and Neuroscience' *Behavioural and Brain Sciences* 30: 481-99.

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

Chalmers, D. and Jackson, F. 2001 “Conceptual Analysis and Reductive Explanation’” *Philosophical Review* 110: 315-61

Chalmers, D. 2002. “Does Conceivability Entail Possibility?” in Hawthorne, J. and Gendler, T. eds *Conceivability and Possibility*. Oxford: Oxford University Press

Davidson, D. 1970 "Mental Events" in Foster, L. and Swanson, J. eds *Experience and Theory* London: Duckworth.

Dennett, D. 1992 *Consciousness Explained*, London: Allen Lane

Feigl, H. 1958 "The 'Mental' and the 'Physical'" in Feigl, H., Scriven, M. and Maxwell, G. eds *Minnesota Studies in the Philosophy of Science vol II* Minneapolis: University of Minnesota Press

Goff, P., 2011 “A Posteriori Physicalists Get Our Phenomenal Concepts Wrong” *Australasian Journal of Philosophy* 89: 191–209.

Goff, P. 2017 *Consciousness and Fundamental Reality* Oxford: Oxford University Press

Howell, R. 2015 “The Russellian Monist’s Problems with Mental Causation”, *Philosophical Quarterly* 65: 22–39

Jackson, F. 1986 “What Mary Didn't Know” *Journal of Philosophy* 83: 291-295.

Jackson, F. 2007 “The Knowledge Argument, Diaphonousness, Representationalism” in Alter, T. and Walter, S. eds *Phenomenal Concepts and Phenomenal Knowledge* Oxford: Oxford University Press

Lee, G. 2014 “Materialism and the Epistemic Significance of Consciousness” in Kriegel, U. ed *Current Controversies in Philosophy of Mind* London: Routledge

Levine, J. 1983 “Materialism and Qualia: The Explanatory Gap” *Pacific Philosophical Quarterly* 64: 354-361

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

Levine, J. 2010 "The Q Factor: Modal Rationalism vs Modal Autonomism" *Philosophical Review* 119: 365-380

Lewis, D. 1966 "An Argument for the Identity Theory" *Journal of Philosophy* 63: 17-25.

Lewis, D. 1988 "What Experience Teaches" *Proceedings of the Russellian Society of Sydney University* 13: 29-57

Lockwood, M. 1993 “The Grain Problem” In Robinson, H. ed., *Objections to Physicalism*, Oxford: Oxford University Press.

McGinn, C. 1991 *The Problem of Consciousness* Oxford: Basil Blackwell.

Menzies, P. 2008 “Causal Exclusion, the Determination Relation, and Contrastive Causation” in Kallestrup, J. and Hohwy, J. eds *Being Reduced: New Essays on Reductive Explanation and Special Science Causation*, Oxford: Oxford University Press.

Milner, A. and Goodale, M. 2008 “Two Visual Systems Reviewed” *Neuropsychologia* 46: 774–85.

Nagel, T. 1986 *The View from Nowhere* Oxford: Oxford University Press

Nida-Rümelin, N. 2007 “Grasping Phenomenal Properties” in Alter, T., and S. Walter eds *Phenomenal Concepts and Phenomenal Knowledge* Oxford: Oxford University Press.

Oppenheim, P. and Putnam, H. 1958 "Unity of Science as a Working Hypothesis" in Feigl, H., Scriven, M. and Maxwell, G. eds *Minnesota Studies in the Philosophy of Science vol II* Minneapolis: University of Minnesota Press.

Papineau, D. 1993 “Physicalism, Consciousness, and the Antipathetic Fallacy” *Australasian Journal of Philosophy* 71:169-83.

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

Papineau, D. 2007 '”Kripke's Argument is Ad Hominem Not Two-Dimensional” in *Philosophical Perspectives* 21: 475-494.

Papineau, D. 2010 "What Exactly is the Explanatory Gap?" *Philosophia* 39: 5-19.

Place, U. 1956 “Is Consciousness a Brain Process?” *British Journal of Psychology* 47: 44–50.

Robinson, W. 2015 “Epiphenomenalism” *Stanford Encyclopedia of Philosophy*

Russell, B. 1927 *The Analysis of Matter* London: Kegan Paul

Shea, N. and Bayne, T. 2010 “The Vegetative State and the Science of Consciousness” *British Journal for the Philosophy of Science* 61: 459-84

Smart, J. 1959 "Sensations and Brain Processes" *Philosophical Review* 68: 141-56

Stoljar, D. 2006 *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness* Oxfird: Oxford University Press

Strawson, G. 2003 “Real Materialism” in Strawson, G. ed *Real Materialism and Other Essays* Oxford: Oxford University Press

Tye, M. 2009 *Consciousness Revisited* Cambridge, Mass: MIT Press

Woodward, J. 2005, *Making Things Happen* Oxford: Oxford University Press