dialectica

dialectica Vol. 67, N° 2 (2013), pp. 117–135
DOI: 10.1111/1746-8361.12016

How to Explain the Explanatory Gap

Neil Mehta[†]

Abstract

I construct a tempting anti-physicalist argument, which sharpens an explanatory gap argument suggested by David Chalmers and Frank Jackson. The argument relies crucially on the premise that there is a deep epistemic asymmetry (which may be identified with the explanatory gap) between phenomenal truths and ordinary macroscopic truths. Many physicalists reject the argument by rejecting this premise. I argue that even if this premise is true, the anti-physicalist conclusion should be rejected, and I provide a detailed, physicalist-friendly explanation of the relevant premise. Along the way, I sketch an account of a priori conceptual knowledge that is compatible with naturalistic accounts of intentionality. I conclude by noting that the resulting view is a version of the popular phenomenal concept strategy that avoids a potentially worrying dilemma facing earlier incarnations of this strategy.

Many philosophers have thought that there is an 'explanatory gap' between the phenomenal and the physical.[1] That is, no matter how much physical information one receives, it seems sensible to ask, "But how could [e.g.] phenomenal redness merely be such-and-such physical property?" This explanatory gap is often taken to show that phenomenal experience cannot be reductively explained in physical terms, so *physicalism* – roughly, the view that everything concrete is reducible to the physical – is false.[2]

This paper focuses on just one tempting version of this argument, the *Explanatory Gap Argument*, which is a sharpened version of an argument suggested by David Chalmers and Frank Jackson – henceforth, CJ – (2001). This argument begins by pointing to a certain epistemic asymmetry, which may be identified with the explanatory gap. The argument then provides a metaphysical explanation of this asymmetry that entails the falsity of physicalism. Many physicalists have responded by denying the existence of the relevant epistemic asymmetry. By contrast, I will show that the Explanatory Gap Argument fails even if we grant the existence of the troublesome asymmetry. The physicalist can provide a perfectly satisfying explanation of the asymmetry – indeed, an explanation that is *clearly superior* to the suggested anti-physicalist explanation. Thus, the physicalist need not quarrel with the claim that there is an explanatory gap of the suggested sort.

[†] Yale-NUS College, Singapore; Email: neil.mehta@yale-nus.edu.sg

[1] See Levine (1983) and (2001).

[2] Nothing in my discussion will hinge on how to characterize physicalism more precisely, but for discussion of this issue, see Crane and Mellor (1990), Jackson (1998), and Lewis (1983).

As it happens, the resulting physicalist-friendly view will have several interesting features. First, it will provide an account of a priori conceptual knowledge that is compatible with thoroughly naturalistic accounts of intentionality. Second, it will be a version of the well-known *phenomenal concept strategy* that smartly sidesteps a potentially worrisome dilemma.[3]

1. The Explanatory Gap Argument

The Explanatory Gap Argument is motivated by the thought that there is an important epistemic difference between phenomenal truths (e.g., *that I am now experiencing phenomenal redness*) and ordinary macroscopic truths (e.g., *that water is transparent*). It seems that we can infer truths about water from truths containing no explicit mention of water. Indeed, it seems that these are *strong inferences*, i.e., inferences that are both conclusive and a priori.[4] By contrast, we cannot strongly infer truths about phenomenal redness from other truths containing no explicit mention of phenomenal redness.

To see how we are supposed to be able to strongly infer macroscopic truths from truths couched wholly in non-macroscopic terms, consider five statements. Let *P* be a truth containing all the fundamental physical information – plausibly, information about the fundamental entities and properties of physics, stated in the language of completed physics. By stipulation, truths about fundamental mental properties and entities, if such there be, are excluded from P. Let *Q* be the conjunction of all phenomenal truths, including truths about what phenomenal properties and states are instantiated at all times. Let *I* be an indexical truth which contains information like 'I am entity X' and 'It is now time Y', where 'X' and 'Y' contain descriptions solely in terms of fundamental physical features of the sort

[3] After independently developing the main ideas of this paper, I came across Joseph Levine's "The Q Factor: Modal Rationalism vs. Modal Autonomism" (2010), which has many points of contact with this paper. It is worth drawing out four substantial differences between this paper and Levine's: (1) Levine's central point is that, on a certain natural reading, the argument of CJ (2001) is circular. I agree, but think that (as Levine also notices) there is another reconstruction of CJ's argument that is not circular. I focus on this version of the argument, whereas Levine discusses it only briefly. (2) Levine and I agree that CJ's argument, properly understood, supports a certain semantic conclusion rather than the metaphysical one that CJ actually draw. But while Levine argues only that CJ provide no reason *to* draw the metaphysical conclusion, I further argue (in section 6) that there are strong reasons *against* drawing this metaphysical conclusion. (3) A central contribution of this paper (in sections 3 to 5) is to develop a detailed physicalist-friendly theory that accommodates the semantic conclusion of CJ's argument. Levine does not pursue this project. (4) Explaining how the physicalist accommodates CJ's semantic conclusion also leads me (again in sections 3 to 5) to a physicalist-friendly account of a priori conceptual knowledge. Again, Levine is not concerned with this matter.

[4] One might think that an inference is conclusive just in case it is deductively valid, or one might think that some non-deductive inferences may also be conclusive. (I thank an anonymous referee for prompting this clarification.)

discussed in P. Let T_D be a statement which says that the world is a *minimal* world satisfying D. Here, a minimal world satisfying D is a world in which D is true and nothing else is true except whatever must be true for D to be true. Finally, let M be the conjunction of all truths about ordinary macroscopic entities and properties. M will include truths like *water is H_2O*, *the earth is smaller than the sun*, and *I am now north of the equator*. By stipulation, M excludes fancy metaphysical claims, normative claims, and mathematical claims.

With these definitions in hand, we can now understand the following pivotal claim:

PQIT-Inferentialism: There is a strong inference from (P & Q & I & $T_{P\&Q\&I}$) to M.

CJ defend this claim vigorously. Consider, for example, the truth that water is H_2O, which is one of the many truths in M. P would likely contain or entail information about the distribution and behavior of H_2O; it would entail that H_2O is located in certain regions and behaves dynamically in certain ways. Combining this information with Q, which contains all the phenomenal information, we could further conclude that H_2O looks, tastes, and feels like water. The indexical truth I would rule out the possibility that, say, water is XYZ rather than H_2O. For this truth would let me figure out that the stuff that causes *me* to have experiences as of water, and which is distributed a certain way in *my* environment, is H_2O, not XYZ. Finally, $T_{P\&Q\&I}$ would rule out the possibility that there is, in addition to H_2O, some other stuff that causes me to have experiences as of water, that fills the rivers and the lakes, etc. It guarantees that H_2O is the *only* entity playing this role. It seems that one can strongly infer from (P & Q & I & $T_{P\&Q\&I}$) to the truth that water is H_2O, since (P & Q & I & $T_{P\&Q\&I}$) permits one to conclusively rule out any hypothesis according to which water is not H_2O.[5] CJ think that similar considerations show that there are strong inferences from (P & Q & I & $T_{P\&Q\&I}$) to any ordinary macroscopic truth whatsoever.

If PQIT-Inferentialism is true, then the following claim looks attractive:

Global Inferentialism: For any global truth E, there is a strong inference from (E & I & $T_{E\&I}$) to M.

Here, a truth is *global* just in case it includes all the information about the concrete metaphysically fundamental entities and properties in the actual world. (A truth that includes more information than that will still count as global.) So if E is a global truth about the actual world, then any minimal world satisfying E will be just like the actual world.

[5] What about someone who lacks some of the concepts required to entertain the claim that if (P & Q & I & $T_{P\&Q\&I}$) is true, then M is true? Such a person is no counterexample to PQIT-Inferentialism. For when considering whether or not some claim is knowable a priori, it is standard to restrict attention to subjects who possess all concepts required to understand that claim.

Global Inferentialism is an appealing explanation of PQIT-Inferentialism. It is quite striking that there is an inference (if there is one) from the physical and phenomenal truths (plus the minor additions of the totality and indexical truths) to M, an enormous conjunction of macroscopic truths; it's very surprising indeed that such an inference is both conclusive and a priori. Global Inferentialism explains why this strong inference would be available. According to Global Inferentialism, PQIT-Inferentialism is just a special case of a more general phenomenon. Quite generally, Global Inferentialism says, there is a strong inference from any global truth E, conjoined with I and $T_{E\&I}$, to M. Since it is quite reasonable to think that the complete truths about the fundamental physical and phenomenal aspects of the world together comprise one such global truth about the world,[6] we get a strong inference from (P & Q & I & $T_{P\&Q\&I}$) to M. One might further suggest that one has a full reductive explanation of some phenomenon just in case one can produce at will any piece of the strong inference from (E & I & $T_{E\&I}$) to that phenomenon.[7]

Let's consider the bearing of Global Inferentialism on physicalism. Physicalists think that P, which contains the complete truth about the fundamental physical aspects of the world, is a global truth. So Global Inferentialism and physicalism together entail this claim:

PIT-Inferentialism: There is a strong inference from (P & I & $T_{P\&I}$) to M.

Is PIT-Inferentialism true? Consider again the truth that water is H_2O, which is one of the many conjuncts of M. It seems clear that some of the phenomenal truths in Q would be necessary to strongly infer this truth, for some of our most fundamental knowledge about water concerns how it phenomenally looks, tastes, etc. So the issue seems to be whether or not there is a strong inference from (P & I & $T_{P\&I}$) to Q. If there is, then given (P & I & $T_{P\&I}$), one can strongly infer to (P & Q & I & $T_{P\&Q\&I}$) and then (if CJ are correct) strongly infer as before to any macroscopic truth M. If not, then a strong inference from (P & I & $T_{P\&I}$) to M will not be available.

Chalmers (1996) and (2009), but not Jackson, further argues that we cannot infer from (P & I & $T_{P\&I}$) to certain phenomenal truths – say, the truth that I am now experiencing phenomenal redness (call this truth Q_R). His argument is that we can conceive, in arbitrarily much detail, of a world in which (P & I & $T_{P\&I}$) is true and Q_R is false. Imagine a *Zombie World* that is just like this one in all respects captured by microphysics, but where there is no phenomenality at all. In this scenario, someone microphysically just like me exists, but he isn't now having any phenomenal experience. But, since this world is just like the actual world in all

[6] Notice that this is compatible with physicalism. If P alone is a global truth, then a fortiori so is (P & Q).

[7] Unsurprisingly, Chalmers (2012, 404–409) expresses sympathy for a thesis much like Global Inferentialism.

fundamental physical respects, (P & I & $T_{P\&I}$) holds (we may suppose that 'I' picks out the appropriate time and individual). If Chalmers is correct that (P & I & $T_{P\&I}$) does not permit a strong inference to Q, then PIT-Inferentialism is false.[8]

If PQIT-Inferentialism is true but PIT-Inferentialism is false, then there is an important epistemic asymmetry between ordinary macroscopic truths and phenomenal truths: ordinary macroscopic truths can be strongly inferred from truths that do not explicitly concern the macroscopic realm, but phenomenal truths cannot be strongly inferred from truths that do not explicitly concern the phenomenal realm. Further, it would not be unreasonable to hold that this epistemic asymmetry just *is* the explanatory gap that generates intuitions of the sort mentioned in the introduction. Thus, for convenience, I will hereafter use the term 'explanatory gap' to refer to this asymmetry. But this is purely a terminological move; those who prefer to use the term 'explanatory gap' in some other way should just treat me as having an unusual way of spelling 'epistemic asymmetry'.

We've arrived at the following anti-physicalist argument, which I dub the *Explanatory Gap Argument*:

(1) PQIT-Inferentialism: There is a strong inference from (P & Q & I & $T_{P\&Q\&I}$) to M. (Supported by examples like 'water is H_2O'.)
(2) Global Inferentialism: For any global truth E, there is a strong inference from (E & I & $T_{E\&I}$) to M. (Best explanation of PQIT-Inferentialism.)
(3) According to physicalism, P is a global truth. (Definition.)
(4) Therefore, if physicalism is true, then PIT-Inferentialism is true: there is a strong inference from (P & I & $T_{P\&I}$) to any macroscopic truth M. (Follows from (2) and (3).)
(5) PIT-Inferentialism is false. (Supported by the Zombie World argument.)
(Conclusion) Therefore, physicalism is false. (Follows from (4) and (5).)

I think that this is the most charitable reconstruction of the arguments presented in CJ's (2001), and I offer textual evidence for this attribution in section 7. However, my primary concern is with the argument itself, whether or not Chalmers or Jackson endorse it.

How are we beleaguered physicalists to respond? Many physicalists respond by denying the alleged epistemic asymmetry – by either denying PQIT-Inferentialism or affirming PIT-Inferentialism.[9] However, I want to bracket such

[8] Chalmers develops additional arguments for the same conclusion, but these arguments are closely related to the Zombie World argument. Notice also that this argument does not require any premise to the effect that conceivability entails possibility. For the conclusion of the Zombie World argument pertains only to *which strong inferences exist*, rather than to *which metaphysical possibilities exist*. I thank an anonymous referee for urging clarification on this point.

[9] Opponents of PQIT-Inferentialism include Block and Stalnaker (1999), Byrne (1999), and Diaz-Leon (2011). Advocates of PIT-Inferentialism include Jackson (2007), Lewis (1966), and Lormand (2004).

moves for present purposes. I will simply grant, for the sake of argument, that this epistemic asymmetry really exists; I will assume that PQIT-Inferentialism is true and that PIT-Inferentialism is false. *Even given these generous concessions*, I will argue, the Explanatory Gap Argument should be resisted, because the metaphysically committal thesis of Global Inferentialism is manifestly *not* the best explanation of PQIT-Inferentialism. There is a clearly superior and purely semantic explanation of PQIT-Inferentialism, which is perfectly compatible with physicalism. Thus, if one identifies this epistemic asymmetry with the explanatory gap, then not only can the physicalist grant the existence of the explanatory gap, but she can even offer a better explanation of this gap than the anti-physicalist.[10]

Here is how I will proceed. Section 2 examines more carefully what it would take for PQIT-Inferentialism to be true. Sections 3 to 5 offer a physicalist-friendly explanation of PQIT-Inferentialism and provide a physicalist-friendly account of a priori conceptual knowledge. Section 6 argues that this explanation is better than the explanation provided by Global Inferentialism. Section 7 provides textual evidence that CJ are committed to the crucial Global Inferentialist premise in the Explanatory Gap Argument. Finally, section 8 concludes by examining the relationship between this view and the popular phenomenal concept strategy.

2. *How could PQIT-Inferentialism be true?*

In this section, I will consider what anyone who accepts PQIT-Inferentialism must be committed to. It's astonishing that anyone could perform the inferences described by PQIT-Inferentialism at all, but how on earth can one make such inferences *conclusively and a priori*?

Think about what I would have to be able to do to strongly infer from (P & Q & I & $T_{P\&Q\&I}$) to M. (P & Q & I & $T_{P\&Q\&I}$) describes the universe using fundamental physical, phenomenal and indexical concepts. It describes the locations (at all times) of all of the fundamental physical particles, forces, fields, etc.; it describes the distribution of all the phenomenal properties and the bearers of those properties; it specifies who one is and what time it is; and, finally, it says that the world is a minimal world satisfying these truths. I can allegedly strongly infer from this description to a description of, say, birds, trees, water, experiences, or what have you. How could I do this?

[10] In fact, I am sympathetic to PQIT-Inferentialism and unsympathetic to PIT-Inferentialism for just the reasons discussed. Still, one who does not share these sympathies and antipathies may still find it interesting that one may comfortably resist anti-physicalism even while making such strong concessions. For this demonstrates the robustness of the physicalist position.

Well, I would have to possess conclusive a priori knowledge of at least some logical and mathematical truths.[11] For P will likely contain mostly descriptions of low-level entities like electrons or quarks, but M will contain mostly statements concerning high-level entities like birds and chairs. I would need to deduce high-level structural properties of entities in the world from P, Q, I, and $T_{P\&Q\&I}$ by performing various logical and mathematical transformations (primarily on P).

But logical and mathematical knowledge would not suffice. No matter how many logical or mathematical transformations one performed on (P & Q & I & $T_{P\&Q\&I}$), one would not get a description in terms of macroscopic entities like birds and chairs. So one would additionally have to know some truths about birds, chairs, etc. to move from the physical-phenomenal-indexical description to M. These truths might be truths like 'If X has a certain shape and size, and is used by person-shaped things to support their weight, and is produced in the right way, and is causally related to me in the right way, and looks enough like entities Y and Z, and . . . , then X is a chair.'

What sorts of truths about macroscopic entities are these? Not empirical truths, for we are considering what can be known a priori about macroscopic entities. The only plausible claim is that these truths which we can know conclusively and a priori are semantic. Moreover, this semantic knowledge does not concern the meanings of our ordinary-language terms. Strong inferences from (P & Q & I & $T_{P\&Q\&I}$) to M are inferences in thought, conducted with concepts; I can make such inferences even if they involve concepts for which I know no corresponding public-language expression. Perhaps I can even make such inferences without knowing any public language at all. So *conceptual* knowledge, of some sort or other, must be the semantic knowledge we are after; conclusive a priori conceptual knowledge must be available if there is a strong inference from (P & Q & I & $T_{P\&Q\&I}$) to M.

Thus, anyone committed to PQIT-Inferentialism must accept two further claims:

> *Analyticity Thesis*: The inference from (P & Q & I & $T_{P\&Q\&I}$) to M requires conceptual truths.
> *Conceptual Apriorism*: We are in a position to know those conceptual truths conclusively and a priori.

Of course, many philosophers would resist these claims, because many philosophers staunchly repudiate conceptual truths. But, for present dialectical purposes, I need not defend the existence of conceptual truths. For this section has argued that one must accept the existence of conceptual truths if one accepts PQIT-Inferentialism, and this paper simply grants the truth of PQIT-Inferentialism for the sake of argument.

[11] What is *conclusive knowledge*? It might simply be *knowledge*, or it might instead be something stronger, like *knowledge grounded in certainty*.

But how might the Analyticity Thesis and Conceptual Apriorism themselves be true? The next section lays the groundwork for a physicalist-friendly response to this question.

3. Basic and non-basic concepts

To motivate this response, let me make some broad remarks about the nature of reference. We can classify any concept according to how it picks out its referent. If the reference relation between a concept and its referent is *wholly mediated by that concept's relations to the subject's other concepts*, then let us call it a *non-basic concept*. Let us call any other concept a *basic concept*. The reference relation between a basic concept and its referent must be *at least partly unmediated by that concept's relations to other concepts* – it must involve at least one direct connection between the basic concept and the referent. Notice, though, that a concept's *non-reference-mediating* relations to other concepts are irrelevant with respect to the basic/non-basic distinction.[12]

To illustrate how a basic concept might work, consider one popular physicalist account of concepts, according to which the reference of at least some concepts is determined purely by their nomological relations to their referents.[13] (As is common, I will use small caps to denote concepts – e.g., DOG is a concept which refers to dogs.) On such accounts, my concept RED (say) refers to redness because red things *normally/ideally cause*, or *cause in a way resulting in asymmetric dependence*, or *have the function of causing*, my application of the concept RED. I'll say that my concept RED *tracks* redness, where 'tracking' is a placeholder for the relevant nomological relation. One plausible suggestion – though certainly not the only possible one – is that any basic concept acquires its reference simply by tracking its referent in a way that is at least partly unmediated by its relations to other concepts.[14] For instance, perhaps my concept RED tracks redness simply because of the non-conceptual activity of subconscious perceptual processes.

Next, consider how a non-basic concept might function. For illustrative purposes only, let's suppose that my concept WATER is non-basic. What might determine the reference of this concept? Well, I possess a vast and frequently updated stock of beliefs about water: that it looks, tastes, and feels a certain way; is the predominant substance in rivers, oceans, and lakes; has an underlying structure that is responsible for its other properties; stands in certain causal relations to me; is called 'water'; fills the glass in front of me; etc. So someone who holds that my

[12] I thank an anonymous referee for very helpful discussion on this point.

[13] See Dretske (1995); Fodor (1990) and (1994); Millikan (1984), (1989), and (2010); Stampe (1977); and Tye (1995) and (2000).

[14] I am in fact sympathetic to this suggestion, but the argument certainly does not require it.

concept WATER is non-basic will hold that it refers to water wholly in virtue of its relationship to some or all of the other concepts implicated in my water-beliefs.

It is a wide-open question precisely how the reference of a non-basic concept is determined. I wish to point out one simple view which might be overlooked, however: the view that any non-basic concept refers by tracking its referent, in just the way that (one might think) any basic concept refers by tracking its referent. My concept WATER, on this view, refers to water because it tracks water. The concept will remain non-basic, however, as long as the tracking relation between my concept WATER and its referent is wholly mediated by the relations between my concept WATER and my other concepts. For instance, perhaps the fact that my concept WATER tracks water is wholly mediated by the fact that it bears certain relations to concepts like WET, RAIN, LAKE, etc.

This brings us to an important point. One can accept the existence of non-basic concepts even while denying that they get their meanings via anything like (stable) *definitions*. For suppose one holds the view just outlined about how my concept WATER refers. Still, one may hold that the very same concept may survive massive change in my beliefs about water, and that there is no particular set of core beliefs that I must hold to maintain the same concept of water. *At any given time*, the fact that my concept WATER refers to water may be mediated by various relations to other concepts, but how this mediation occurs may vary dramatically *across times*.

Stepping back a bit, the distinction between basic and non-basic concepts suggests the following picture. Basic concepts get their reference at least partly by hooking up directly to their referents in some way or other – plausibly, via tracking. Non-basic concepts get their reference indirectly, via their relations to other concepts; those other concepts might in turn be basic or non-basic. Any of these other non-basic concepts will themselves get their reference via further basic or non-basic concepts. But, ultimately, this process must bottom out in basic concepts – it cannot be non-basic concepts all the way down.[15] Thus, fixing the reference of all of one's basic concepts will suffice to fix the reference of one's non-basic concepts, too.

4. A minimal hypothesis, part 1

Now for the first of my two hypotheses to explain the truth of the Analyticity Thesis and Conceptual Apriorism. Call the concepts we use to entertain (P & Q &

[15] It *could*, however, be basic concepts all the way down. Though many philosophers – a representative example being Fodor (1998) – think that all concepts are basic, such philosophers invariably deny PQIT-Inferentialism, too. Since we're examining what it would take for PQIT-Inferentialism to be true, we must grant that there are some non-basic concepts.
It could also be that some sets of concepts mutually constrain one another's reference. Perhaps WATER gets its reference partially via its relation to WET, and WET also gets its reference partly via its relation to WATER. Nothing I have said is incompatible with this possibility.

I & $T_{P\&Q\&I}$) our *PQIT concepts*, and call the concepts we use to entertain the macroscopic truth M our *macroscopic concepts*. I propose the following hypothesis:

> *PQIT-Determination*: Fixing the reference of our PQIT concepts suffices to fix the reference of all of our macroscopic concepts.

Note that PQIT-Determination entails that it is a conceptual truth, in some good sense of that phrase, that if (P & Q & I & $T_{P\&Q\&I}$) holds, then so does M. Assuming that the strong inference from (P & Q & I & $T_{P\&Q\&I}$) to M relies on such conceptual truths, PQIT-Determination entails the Analyticity Thesis.

Let me explain in some detail how PQIT-Determination could be true. Plausibly, there are basic macroscopic concepts (perhaps SPACE and CAUSE) and non-basic macroscopic concepts (perhaps WATER and WET). Similarly, there are basic PQIT concepts (perhaps PHENOMENAL REDNESS and NOW) and non-basic PQIT concepts (perhaps PROTON or CHARGE). Moreover, there will be some overlap between our macroscopic concepts and our PQIT concepts – examples might include SPACE and CAUSE.

I will first argue, by *reductio*, that any basic macroscopic concept must also be a PQIT concept. Suppose, for *reductio*, that there is some basic macroscopic concept (say, the concept DOG) that is *not* a PQIT concept. Now, how can one strongly infer truths about dogs? This strong inference cannot rely on conceptual truths linking (P & Q & I & $T_{P\&Q\&I}$) to dogs. For since the concept DOG is hypothesized to be a basic concept, there are no such conceptual truths. Further, it is utterly mysterious how else we might strongly infer truths about dogs from (P & Q & I & $T_{P\&Q\&I}$). So there can be no strong inference from (P & Q & I & $T_{P\&Q\&I}$) to all truths about dogs. This contradicts PQIT-Inferentialism, so our *reductio* assumption must have been mistaken. Therefore, all basic macroscopic concepts are PQIT concepts. Accordingly, fixing the reference of our PQIT concepts will trivially fix the reference of all basic macroscopic concepts.[16]

Next, I will argue that all non-basic macroscopic concepts get their reference via the PQIT concepts. This argument is similar to CJ's argument that one can strongly infer from (P & Q & I & $T_{P\&Q\&I}$) to M. Consider some macroscopic concepts – say, WATER, BIRD and GOLD. Like most macroscopic concepts, these concepts are intimately associated with the phenomenal information contained in Q: I recognize water, birds, and gold largely by how they look, feel, etc. Of course, such information is insufficient to determine the reference of my concepts WATER, BIRD and GOLD, for several reasons.

First, I may conceive of water, birds, and gold as having a certain hidden structure. Information about the shape, motion, and behavior of microphysical particles provided by P will pin down this hidden structure. Second, I conceive of

[16] I thank an anonymous referee for constructive comments on this point.

water, birds, and gold as being causally related to me in certain ways. The indexical truth I will provide this kind of information.

Third, I pick out some of these entities (e.g., birds) partly via their macroscopic properties (like having wings, having a certain ancestral history, or being called 'birds' by the experts). So to determine whether or not something is a bird, I may consider whether or not it has these properties. But, ultimately, non-basic concepts like WINGS, OFFSPRING and EXPERTS must have their reference determined via their relations to basic concepts. And, as before, the physical, phenomenal and indexical concepts look like the best candidates to do this. So it is reasonable to think that the PQIT concepts determine the reference of all macroscopic concepts, just as PQIT-Determination says.

PQIT-Determination gets further support from the fact that the Analyticity Thesis must be true (assuming, as we are, that PQIT-Inferentialism is true). It must be some kind of conceptual truth that if (P & Q & I & $T_{P\&Q\&I}$) holds, then so does M. It is plausible that this is a conceptual truth because the basic PQIT concepts include the basic macroscopic concepts, and the PQIT concepts together determine the reference of the non-basic macroscopic concepts.

5. *A minimal hypothesis, part 2*

Here's my second key hypothesis for explaining PQIT-Inferentialism:

> *Restricted Conceptual Apriorism*: If fixing the reference of some set of concepts S_1 suffices to fix the reference of some set of concepts S_2, then, if I am given the reference of the concepts in S_1, then I can determine the reference of the concepts in S_2 conclusively and a priori.

Let me illustrate what Restricted Conceptual Apriorism says. Suppose that the reference of my concept WATER is wholly mediated by some or all of the concepts implicated in my water-beliefs – e.g., the beliefs that water looks and tastes a certain way, is the predominant substance in the rivers and lakes, quenches thirst, etc. Then, according to Restricted Conceptual Apriorism, given the information that H_2O looks and tastes that way, is the predominant substance in the rivers and lakes, and quenches thirst, I can determine, conclusively and a priori, that water is H_2O.

Why think that Restricted Conceptual Apriorism is true? To begin with, I argued in section 2 that someone who accepts PQIT-Inferentialism is committed to a priori conceptual knowledge, and Restricted Conceptual Apriorism posits a relatively weak brand of such knowledge of just the sort required to capture PQIT-Inferentialism. Further, Restricted Conceptual Apriorism would help explain our ability to pick out the referents of our concepts in various epistemic possibilities. For example, I can imagine discovering that I inhabit Twin Earth,

where XYZ looks and tastes a certain way, is the predominant substance in the rivers and lakes, quenches thirst, etc. I know that if this possibility actually obtains, then XYZ is water. Restricted Conceptual Apriorism explains how I acquire such knowledge: I take for granted certain claims about the reference of the concepts which wholly mediate the reference of the concept WATER, and this puts me in a position to determine the reference of the concept WATER itself.

Together, PQIT-Determination and Restricted Conceptual Apriorism straightforwardly explain PQIT-Inferentialism. According to PQIT-Determination, fixing the reference of the concepts that we use to entertain (P & Q & I & $T_{P\&Q\&I}$) suffices to fix the reference of all concepts required to entertain M; put another way, it is a conceptual truth that, if (P & Q & I & $T_{P\&Q\&I}$) holds, then M holds. According to Restricted Conceptual Apriorism, one can figure out the reference of a concept conclusively and a priori once one knows the reference of all the concepts which suffice to fix its reference. Combining these claims, we arrive at PQIT-Inferentialism, the claim that there is a strong inference from (P & Q & I & $T_{P\&Q\&I}$) to macroscopic truth M. Call this the *Semantic Explanation* of PQIT-Inferentialism.

(Incidentally, this also suggests an explanation of the falsity of PIT-Inferentialism. If PQIT-Inferentialism is true but PIT-Inferentialism is false, then one natural explanation would be that some concepts used to entertain Q are basic. For if some concepts used to entertain Q are basic, then there will be no conceptual truth linking (P & I & $T_{P\&I}$) to Q, which would explain the lack of a strong inference from (P & I & $T_{P\&I}$) to Q. If we use the term *phenomenal concept* to refer to any concept distinctively used to think about phenomenality, then the idea is that some phenomenal concepts are basic.)

Notably, the Semantic Explanation is perfectly compatible with either physicalism or non-physicalism. For PQIT-Determination says only that certain reference-fixing relations obtain, and Restricted Conceptual Apriorism says only that we have the ability to determine the reference of some of our concepts under certain conditions. These purely semantic/epistemic claims do not immediately entail any metaphysical consequences. Indeed, the next section will show that the physicalist can escape the Explanatory Gap Argument by invoking the Semantic Explanation.

6. Evaluating the explanations

The Semantic Explanation purports to explain PQIT-Inferentialism; so does Global Inferentialism, which says that for any global truth E, there is a strong inference from (E & I & $T_{E\&I}$) to M. The Semantic Explanation, as the name suggests, explains PQIT-Inferentialism in terms of the *semantic* distinction between basic and non-basic concepts. Global Inferentialism explains PQIT-

Inferentialism *metaphysically*, by positing strong inferences wherever there are global truths. What is the relationship between these two explanations?

The Global Inferentialist must accept either the Semantic Explanation or something very much like it. For, as I argued in section 2, *anyone* who accepts PQIT-Inferentialism will have to accept both the Analyticity Thesis and Conceptual Apriorism. The Semantic Explanation is the conjunction of PQIT-Determination and Restricted Conceptual Apriorism, which are more specific versions of the Analyticity Thesis and Conceptual Apriorism. Even if the Global Inferentialist does not like these specific formulations, she will have to settle for something in the vicinity.

But one can accept the Semantic Explanation while rejecting Global Inferentialism. On this view, PQIT-Inferentialism is true because fixing the reference of our PQIT-concepts suffices to fix the reference of our macroscopic concepts. But this semantic point is separate from metaphysics; it indicates nothing about the availability of strong inferences from global truths more generally. Call this the *Minimal Semantic Explanation*.

The heart of the matter, then, is whether one should prefer Global Inferentialism or the Minimal Semantic Explanation. I now offer (in increasing order of dialectical forcefulness) three reasons for favoring the Minimal Semantic Explanation.

First, Global Inferentialism (but not the Minimal Semantic Explanation) will drive one towards dualism, as the Explanatory Gap Argument shows. This is a prima facie cost. Even dualists should admit that dualism is a prima facie worse theory than physicalism, as it is more ontologically inflationary. (Dualists merely think that dualism compensates for this lack of parsimony by better handling certain data about phenomenal experience.)

Second, and more tellingly, the extra commitments of Global Inferentialism are explanatorily idle. The Global Inferentialist and the Minimal Semantic Explanation theorist both accept the Semantic Explanation, but the Global Inferentialist further draws a connection between semantics and metaphysics: she says that for any *global truth* E, there is a strong inference from (E & I & $T_{E\&I}$) to M. (The 'global truth' bit is where Global Inferentialism makes a metaphysical commitment.) The crucial point, though, is that the putative semantics-metaphysics connection plays no additional role in explaining PQIT-Inferentialism. The Semantic Explanation alone does all the heavy lifting; as long as it is correct, PQIT-Inferentialism will hold, *regardless* of whether there is any semantics-metaphysics connection. So the additional metaphysical commitments of the Global Inferentialist are, at least thus far, unmotivated.

Here is the third, and most damning, reason for preferring the Minimal Semantic Explanation to Global Inferentialism. Consider this claim, which is very similar to PQIT-Inferentialism:

PQT-Inferentialism: There is a strong inference from (P & Q & $T_{P\&Q}$) to M.

PQT-Inferentialism is demonstrably false. For, as is well known, there is no way to strongly infer indexical truths (like 'I am here') from truths couched in non-indexical terms.[17] Moreover, it is unlikely that there is a strong inference from (P & Q & $T_{P\&Q}$) even to non-indexical truths like 'water is H_2O'. For I conceive of water as the stuff of *my* acquaintance which has certain features, and without indexical information about who I am, I cannot strongly infer which stuff is the right stuff of *my* acquaintance.

The Global Inferentialist has great difficulty explaining why PQT-Inferentialism is false while PQIT-Inferentialism is true. Put crudely, the picture that the Global Inferentialist would like to endorse is that any global truth permits a strong inference to any truth whatsoever.[18] Moreover, (P & Q & $T_{P\&Q}$) is a global truth. But there is not a strong inference from (P & Q & $T_{P\&Q}$) alone to M; the indexical truth I is required as well. The Global Inferentialist has no obvious explanation of why *this* particular truth about non-fundamental matters should be necessary for strongly inferring M. She endorses PQIT-Inferentialism rather than PQT-Inferentialism because the latter is extensionally inadequate, but she has no further rationale for this shift. Tellingly, all that CJ say in defense of the shift from PQT-Inferentialism to PQIT-Inferentialism is that the addition of I is 'sufficiently minor' and "[does] not change much" (Chalmers and Jackson 2001, 351).[19]

By contrast, the Minimal Semantic Explanation theorist can easily explain why PQT-Inferentialism is false but PQIT-Inferentialism is true. The key claim is that we possess some basic indexical concepts. Because these concepts (like some phenomenal concepts) are basic, we need not be able to strongly infer from (P & Q & $T_{P\&Q}$) to claims featuring such indexical concepts. The fact that (P & Q & $T_{P\&Q}$) is a global truth is irrelevant.[20]

I conclude that the physicalist-friendly Minimal Semantic Explanation is clearly – indeed, almost *strictly* – superior to the physicalist-unfriendly Global Inferentialist explanation. The Explanatory Gap Argument can be successfully

[17] See Lewis (1979) and Perry (1979).

[18] See Chalmers (2012, 404–409).

[19] An anonymous referee suggested that Global Inferentialism might have the following crucial advantage over the Minimal Semantic Explanation. On both views, there is something special about certain concepts, including certain phenomenal, physical and indexical concepts. But the Global Inferentialist has a tidy account of this specialness: such concepts refer to the metaphysically fundamental furniture of the universe. However, I have just argued the Global Inferentialist's claim is simply false: at least some indexical concepts don't refer to any metaphysically fundamental furniture of the universe. Thus, any 'advantage' that the Global Inferentialist has here is spurious.

[20] If indexical concepts like I and NOW are basic, how might they get their reference? Perhaps at least partly via their tendency to trigger action directly. If I believe that the pants of the person in the mirror are on fire, I may not do anything. But if I believe that *my* pants are on fire, I will immediately try to extinguish the flames. See Millikan (1990) for a related suggestion.

resisted even if we concede the existence if a certain sort of explanatory gap – that is, even if we concede the truth of PQIT-Inferentialism and the falsity of PIT-Inferentialism. Better yet, in the course of showing this, we have drawn up a detailed blueprint for a physicalist-friendly explanation of this explanatory gap.

7. Are CJ committed to Global Inferentialism?

CJ do not explicitly state Global Inferentialism anywhere in their (2001). Still, as I will argue in this section, they are committed to this claim. (Readers uninterested in Chalmers-Jackson exegesis should bypass this section.) The first reason for thinking this is that the best anti-physicalist argument I can extract from CJ's paper, the Explanatory Gap Argument, requires Global Inferentialism. I leave an open invitation to CJ (or the reader) to provide a better anti-physicalist argument from the considerations raised in that paper.

More importantly, there is abundant textual evidence that CJ accept Global Inferentialism. CJ say this in the opening lines of their (2001): "If there is no [strong inference] from microphysical truths to phenomenal truths, does reductive explanation of the phenomenal fail? We say yes." Why think this? Presumably because one thinks, quite generally, that something like the following holds:

> *Ultrastrong Local Inferentialism*: If truth E* is fully reductively explainable in terms of truth E, then there is a strong inference from E to E*.

Ultrastrong Local Inferentialism can't be quite what Chalmers accepts (the considerations I raise won't apply to Jackson). For consider the truth 'water is liquid at room temperature'. This truth is reductively explainable in terms of the physical truths alone – phenomenal truths aren't any part of the reductive explanation of water's liquidity at room temperature. However, there is no strong inference from the physical truths alone to the truth 'water is liquid at room temperature'. For we pick out water largely via its phenomenal effects on us, and the physical information alone won't inform us of these effects. (Recall that Chalmers thinks that the physical truths alone don't a priori entail the phenomenal truths.)

CJ aren't after a principle about which truths reduce to which other *local* truths. Rather, they're after a principle about which *global* truth all other truths reduce to. So here's a plausible revision of Ultrastrong Local Inferentialism:

> *Ultrastrong Global Inferentialism*: There is a strong inference from any global truth E to any truth at all.

But CJ (2001, 335–336) back off this strong claim:

> we have not argued that (P & Q & I & $T_{P\&Q\&I}$) implies every truth in every domain (though we are inclined to accept this claim). Given what we have said here, it could be that certain truths in special domains – perhaps concerning mathematics, metaphysics, morality, or mentality? – are not implied by P & Q & I & $T_{P\&Q\&I}$) . . . But we

> hope we have said enough to make it plausible that ordinary macroscopic truths concerning everyday macroscopic natural phenomena [may be strongly inferred from] (P & Q & I & $T_{P\&Q\&I}$).[21]

This suggests a retreat to this slightly weaker claim:

> *Strong Inferentialism*: There is a strong inference from any global truth E to any macroscopic truth M.

And the following passage suggests that CJ would like to accept Strong Inferentialism:

> When a concept of some natural phenomenon supports a priori entailments from the microphysical, there is a clear sense in which the phenomenon can be reductively explained. These a priori entailments might not support a *reduction* of the phenomenon in question to a microphysical phenomenon (at least in some senses of this term), perhaps because such entailments are compatible with multiple realizability. But, nevertheless, in showing how any instance of the phenomenon is itself implied by microphysical phenomena, we show that there is a sort of transparent epistemic connection between the microphysical and macroscopic phenomena. (Ib., 350–351)

The considerations raised in this passage seem to support something like Strong Inferentialism. Even Strong Inferentialism is too strong, though. For physicalists and anti-physicalists alike can agree that (P & Q) is a global truth (though physicalists also think that P alone is a global truth). However, (P & Q) alone won't support strong inferences to all of macroscopic truths. For, as was demonstrated by the earlier example concerning the truth 'water is H_2O', many such strong inferences would require knowledge of the indexical truth I and totality truth T, neither of which is strongly inferable from (P & Q). CJ recognize this problem:

> When a phenomenon is entailed a priori by (P & Q & I & $T_{P\&Q\&I}$) . . ., something similar applies . . . T and I are sufficiently minor additions to the reduction base that they do not change much. (Ib., 351)

Taking this suggestion into account, we arrive at the thesis which, so far as I can tell, CJ accept without qualification:

> *Global Inferentialism*: For any global truth E, there is a strong inference from (E & I & $T_{E\&I}$) to any macroscopic truth M.

8. *Conclusion*

I began this paper by granting my opponents the truth of PQIT-Inferentialism and the falsity of PIT-Inferentialism, an epistemic asymmetry which might reasonably be construed as the explanatory gap. I have argued that the (physicalist-friendly) Minimal Semantic Explanation of this explanatory gap is clearly superior to the

[21] CJ use a slightly different notation in place of "(P & Q & I & $T_{P\&Q\&I}$)"; for consistency, I replace their notation with mine even when quoting them.

(physicalist-unfriendly) Global Inferentialist explanation. Thus, this construal of the explanatory gap has no obvious anti-physicalist import.

I want to conclude by briefly situating the resulting view in a broader context. The present view can fruitfully be understood as a version of the popular *phenomenal concept strategy*, introduced by Loar (1990/1997).[22] The broad idea behind the phenomenal concept strategy is to explain away anti-physicalist data – in this case, the explanatory gap, construed in a certain way – in terms of special semantic features of phenomenal concepts, rather than special metaphysical features of phenomenal consciousness. That is precisely what the Minimal Semantic Explanation does: it explains the explanatory gap by exploiting the special semantic features of basic concepts. The phenomenal concept strategy is familiar and well-explored, but let me briefly draw attention to a distinctive advantage of this implementation of the phenomenal concept strategy.

Previous phenomenal concept strategists have typically fallen into one of two camps. The first camp of phenomenal concept strategists hold that phenomenal concepts are *sui generis*: they have special features that set them apart from all other kinds of concepts, and those special features explain the relevant data. At worst, this approach invites the charge of being *ad hoc*; at best, it takes on the additional burden of explaining why phenomenal concepts are special in this way. Perhaps this burden can ultimately be discharged, but it would certainly be nice to avoid shouldering the burden to begin with.

This brings us to the second camp of phenomenal concept strategists, who hold that phenomenal concepts are deeply similar to some other class of concepts – say, recognitional concepts,[23] or concepts that are, broadly speaking, indexical.[24] Such accounts shoulder a different burden: they must show that phenomenal concepts function like these other concepts in a fine-grained way. However, some have responded that the putative analogies between phenomenal concepts and these other sorts of concepts rapidly break down.[25] Without closely examining the merits of these complaints, we can note that it would surely be nice to have a view which can account for the data without relying on detailed commitments about the fine-grained semantic features of phenomenal concepts.

Thus, the phenomenal concept strategist faces this (admittedly soft) dilemma: if phenomenal concepts are *sui generis*, then we must explain what is special about them; if not, then we must accept detailed and highly contentious commitments about the fine-grained semantic features of phenomenal concepts. Notably,

[22] For further discussion of the phenomenal concept strategy, see Alter and Walter (2007).

[23] See Loar (1990/1997).

[24] This approach has been *very* popular; representative examples include Lycan (1996), Papineau (2007), and Perry (2001).

[25] See Chalmers (2004).

however, Minimal Semantic Explanation elegantly avoids both horns of this dilemma. On this view, phenomenal concepts are basic concepts, and there are many non-phenomenal basic concepts. So, in a broad sense, phenomenal concepts are not *sui generis*, and the first horn of the dilemma is avoided. However, this requires no detailed commitments about the semantic features of phenomenal concepts. For to hold that some class of concepts is basic is to make a primarily negative claim: that the reference of each concept of this class is *not* wholly mediated by its relation to other concepts. This claim provides no positive story about how the reference of any basic concept *is* mediated, and indeed leaves open the possibility of great semantic heterogeneity across different classes of basic concepts. Accordingly, in a narrow and unproblematic sense, phenomenal concepts may well be *sui generis*.

Thus, not only is the Minimal Semantic Explanation a physicalist-friendly explanation of the explanatory gap, but it just might also be the most promising version of the phenomenal concept strategy. But pursuing this claim further is a project for another occasion.*

References

Alter, T. and Walter, S. 2007, *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, New York: Oxford University Press.

Block, N. and Stalnaker, R. 1999, "Conceptual Analysis, Dualism, and the Explanatory Gap", *Philosophical Review*, **108**, pp. 1–46, reprinted in: D. Chalmers, ed., *Philosophy of Mind: Classical and Contemporary Readings*, New York: Oxford University Press, 2002, pp. 371–394.

Byrne, A. 1999, "Cosmic Hermeneutics", *Philosophical Perspectives*, **13**, pp. 347–383.

Chalmers, D. 1996, *The Conscious Mind*, New York: Oxford University Press.

Chalmers, D. 2004, "Imagination, Indexicality, Intensions", *Philosophy and Phenomenological Research*, **68**, pp. 182–190.

Chalmers, D. 2009, "The Two-Dimensional Argument Against Materialism", in: B. McLaughlin, ed., *The Oxford Handbook of the Philosophy of Mind*, Oxford: Oxford University Press, pp. 313–335.

Chalmers, D. 2012, *Constructing the World*, New York: Oxford University Press.

Chalmers, D. and Jackson, F. 2001, "Conceptual Analysis and Reductive Explanation", *Philosophical Review*, **110**, pp. 315–361.

Crane, T. and Mellor, D. 1990, "There is No Question of Physicalism", *Mind*, **99**, pp. 185–206.

Diaz-Leon, E. 2011, "Reductive Explanation, Concepts, and A Priori Entailment", *Philosophical Studies*, **155**, pp. 99–116.

Dretske, F. 1995, *Naturalizing the Mind*, Cambridge, MA: MIT Press.

Fodor, J. 1990, *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press.

Fodor, J. 1994, *The Elm and the Expert*, Cambridge, MA: MIT Press.

Fodor, J. 1998, *Concepts: Where Cognitive Science Went Wrong*, New York: Oxford University Press.

* I am especially indebted to Eric Lormand for his tireless readings of numerous drafts of this paper. I also received helpful feedback from David Braddon-Mitchell; Victor Caston; Andy Egan; Jens Kipper; Andrew McGonigal; participants in the 2012 University of Geneva workshop "What did we learn from two-dimensionalism?"; a 2008 meeting of the Australian National University Philosophy Society; the 2008 candidacy seminar at the University of Michigan; and several anonymous referees for *dialectica*.

Jackson, F. 1998, *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford: Clarendon Press.

Jackson, F. 2007, "A Priori Physicalism", in: B. McLaughlin and J. Cohen, eds, *Contemporary Debates in Philosophy of Mind*, Malden, MA: Blackwell Publishing, pp. 185–199

Levine, J. 1983, "Materialism and Qualia: The Explanatory Gap", *Pacific Philosophical Quarterly*, **64**, pp. 354–361.

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

Levine, J. 2010, "The Q Factor: Modal Rationalism vs. Modal Autonomism", *Philosophical Review*, **119**, 3, pp. 365–380.

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

Lewis, D. 1979, "Attitudes *de dicto* and *de se*", *Philosophical Review*, **88**, pp. 513–543.

Lewis, D. 1983, "New Work for a Theory of Universals", *Australasian Journal of Philosophy*, **61**, 4, pp. 343–377.

Loar, B. 1990/1997, "Phenomenal States," *Philosophical Perspectives*, **4**, pp. 81–108, revised version in: N. Block, O. Flanagan and G. Güzeldere, eds, *The Nature of Consciousness*, Cambridge, MA: MIT Press, 1997, pp. 597–616.

Lormand, E. 2004, "The Explanatory Stopgap", *Philosophical Review*, **113**, pp. 303–357.

Lycan, W. 1996, *Consciousness and Experience*, Cambridge, MA: MIT Press.

Millikan, R. 1984, *Language, Thought and Other Biological Categories*, Cambridge, MA: MIT Press.

Millikan, R. 1989, "Biosemantics", *Journal of Philosophy*, **86**, pp. 281–297.

Millikan, R. 1990, "The Myth of the Essential Indexical", *Noûs*, **24**, pp. 723–734.

Millikan, R. 2010, "On Knowing the Meaning: With a Coda on Swampman", *Mind*, **119**, pp. 43–81.

Papineau, D. 2007, "Phenomenal and Perceptual Concepts", in: T. Alter and S. Walter, eds, *Phenomenal Concepts and Phenomenal Knowledge*, New York: Oxford University Press, pp. 111–144.

Perry, J. 1979, "The Problem of the Essential Indexical", *Noûs*, **13**, pp. 3–21.

Perry, J. 2001, *Knowledge, Possibility, and Consciousness*, Cambridge, MA: MIT Press.

Stampe, D. 1977, "Towards A Causal Theory of Linguistic Representation", *Midwest Studies in Philosophy*, **2**, 1, pp. 42–63.

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

Tye, M. 2000, *Consciousness, Color, and Content*, Cambridge, MA: MIT Press.