Dissolving Type-B Physicalism

Helen Yetter-Chappell[1]

The majority of physicalists are type-B physicalists – believing that the phenomenal-physical truths are only knowable a posteriori. This paper aims to show why this view is misguided. The strategy is to design an agent who (1) has full general physical knowledge, (2) has phenomenal concepts, and yet (3) is wired such that she would be in a position to immediately work out the phenomenal-physical truths. I argue that this derivation yields a priori knowledge. The possibility of such a creature entails that – contrary to type-B physicalism – there is not an *ideal* epistemic gap between the phenomenal and the physical truths. Out of this argument against type-B physicalism emerges a positive result: a new and compelling version of type-A physicalism, roughly a type-A phenomenal concept strategy.

§1 Introduction

A significant majority of physicalists believe that phenomenal-physical truths are only knowable a posteriori – that zombies are ideally conceivable, but not possible.[2] This paper explains why these "type-B" physicalists are misguided. If physicalism is true, we can design an agent with a priori access to the phenomenal-physical truths. I show how to design such an agent.

While many philosophers have argued that physicalists are committed to the "type-A" view that phenomenal-physical truths are knowable a priori (Chalmers 1999, Strawson 2006), these arguments have depended on principles – that the phenomenal-physical truths must be "intelligible to God" or that there are no strong necessities – which are contrary to the core commitments of type-B physicalism. By contrast, this paper's aim is to provide an argument against type-B physicalism on grounds that its proponents should accept.

---

[1] Thanks to Torin Alter, David Chalmers, Mark Harris, Terry Horgan, Frank Jackson, Sarah-Jane Leslie, Bill Lycan, Corey Maley, Jimmy Martin, Angela Mendelovici, Kristin Primus, Derek Shiller, Galen Strawson, and Jack Woods for their feedback on various drafts of this paper. And special thanks to Richard Yetter Chappell, for all his helpful discussions of the ideas in this paper and for his feedback across numerous drafts.

[2] From the 2009 philpapers survey of "target philosophers" (philpapers.org/surveys): Among the 138 philosophers of mind who either take zombies to be "conceivable but not metaphysically possible" or "inconceivable", 66% hold that they are "conceivable but not metaphysically possible".

To bring out the distinction between type-B and type-A physicalism, consider how these views respond to anti-physicalist arguments. These arguments center around the idea that there's an ideal epistemic and explanatory gap between the physical truths (P) and the phenomenal truths (Q), from which an ontological gap can be inferred. It seems a fully informed, fully rational agent could conceive of a zombie world: a world physically just like our world, but with no phenomenal experiences (a world where P & ~Q is true). Mary, the brilliant color scientist who's been raised in a black and white room and never seen red, knows all the physical facts, but she doesn't know what red experiences are like. Mary knows P and is fully rational, but fails to know what red experiences are like (Q). Further, there seems to be an explanatory gap between P and Q. When we consider *why*, given that P is the case, Q is the case, there seems to be no answer – and we seem to need an answer! It seems completely arbitrary that the neurological state I'm in feels the way it does or feels like anything at all. If the epistemic/explanatory gaps entail a metaphysical gap, then it is metaphysically possible for P to obtain without Q, in which case physicalism is false.

Type-A physicalists deny that there is an ideal epistemic or explanatory gap. We are misled into believing there are such gaps, either because our own physical knowledge is incomplete (and phenomenal truths don't follow a priori from an incomplete set of physical truths) or because our rational capacities are limited. Type-B physicalists grant that there are ideal epistemic and explanatory gaps between the physical truths and the phenomenal truths but deny that this entails an ontological gap. The type-B physicalist seeks to accommodate our dualist intuitions without accepting a dualist ontology.

A simple defense of type-B physicalism might appeal to uncontroversial cases of necessary a posteriori truths: "It's conceivable that water not be H_2O, though this is not metaphysically possible. So the move from ideal conceivability to possibility is misguided. We can embrace our intuitions and have physicalism too." Unfortunately, the physicalist can't model her account of the necessary a posteriori status of phenomenal-physical truths on these uncontroversial cases – the anti-physicalist arguments are powerful precisely because they bring out points of disanalogy with standard cases of the necessary a posteriori. While I may be able to conceive of a world in which water is not H_2O, if I knew all the microphysical facts, I would not be able to imagine this (compatible with my actual knowledge). By

contrast, it seems that I could know all of the physical facts and yet still be able to conceive of a zombie world (compatible with my actual knowledge).[3] And, although there is an epistemic gap between water and H_2O, there is no explanatory gap. Once we learn that water is H_2O, we are not left with a nagging feeling that the identity is somehow arbitrary or that they can't possibly be one and the same thing. But we do continue to wonder these things about phenomenal-physical identities.

If the necessary a posteriori status of phenomenal-physical truths cannot be explained in the familiar way, type-B physicalists must give a novel account of the novel epistemic and explanatory gap that we find in the case of phenomenal-physical truths. The Phenomenal Concept Strategy (PCS) has emerged as an attractive and popular way to do this.[4]

According to PCS, phenomenal concepts are so radically different from physical concepts that even an ideally rational agent with all of the relevant concepts and complete physical information could not work out the phenomenal-physical truths a priori. (Different theories of phenomenal concepts explain this radical conceptual disconnect in different ways.) Suppose that a physicalistically acceptable account of phenomenal concepts can be given that vindicates this move. While this would not *close* the epistemic and explanatory gaps, it does the next best thing: It gives a physicalistically respectable story of why the gaps arise.

This paper shows that no version of the phenomenal concept strategy can succeed. I show that for any theory of phenomenal concepts, we can construct an agent to serve as a counterexample to the claim that an ideal agent would find an epistemic gap. If right, this establishes the falsity of type-B physicalism more generally, by offering a concrete demonstration of how the phenomenal-physical truths can be known a priori.

My strategy is to design an agent who (1) has full general physical knowledge, (2) has phenomenal concepts that work just as your favored theory of phenomenal concepts claims, and yet (3) is wired such that she would be in a position to immediately work out the phenomenal-physical truths. I argue that – if physicalism is true – this creature's immediate phenomenal-physical judgments amount to a priori knowledge. It follows from this that there

---

[3] See: Chalmers (1999), Jackson (2000), Chalmers & Jackson (2001)

[4] See, for example: Loar (1990/1997), Lycan (1996), Perry (2001), Papineau (2002, 2007), Tye (2003), Levin (2007), Balog (2012).

is *not* an ideal epistemic gap between the phenomenal truths and the physical truths.

1.1 Implications

Type-B physicalism is the most popular version of physicalism. If my arguments are right, this view is untenable. This is a striking and important result.

Perhaps equally important is the positive view that emerges from the failure of the phenomenal concept strategy: a new and highly compelling way to be a type-A physicalist. *We* clearly have a very different psychology from the imaginary agent I'll describe. I suggest that it is this psychological difference of ours – we might say "this rational defect" of ours – that explains why *we* find an epistemic and explanatory gap. I argue that this gets the type-B physicalist what they really wanted all along: a way to be a physicalist that respects the intuitions that *Mary* would be fooled if her captors presented her with a blue banana, that *we* would find zombies conceivable no matter how much physical information we possessed, that *we* will never be able to bridge the explanatory gap. Accepting this version of physicalism requires abandoning the idea that these gaps are *ideal*, but plausibly our intuitions about zombie cases and the like were never about what creatures with radically different psychologies from our own would be able to conceive of; rather they were about creatures psychologically like us.

This should be attractive to physicalists of all varieties: Type-A physicalists can continue to maintain a tight connection between ideal conceivability and possibility, while respecting our most firmly held intuitions (something they've had difficulty doing). And though my argument shows that the letter of type-B physicalism is untenable – physicalism is not compatible with an ideal epistemic gap – we have a new way to be a physicalist that captures what I take to be the *spirit* of type-B physicalism. We have a way to be a physicalist while respecting our dualistic intuitions, avoiding messy debates about the relationship between conceivability and possibility.

1.2 A Roadmap

I argue that, given any plausible theory of phenomenal concepts, we can construct an

agent who is able to determine the phenomenal-physical truths a priori. I begin in §2 with what I take to be the most plausible theory of phenomenal concepts: the constitutional theory. I sketch a psychology that would enable an agent who possessed constitutional phenomenal concepts to work out the phenomenal-physical truths directly. I then argue that these immediate phenomenal-physical judgments would amount to a priori knowledge. §3 defends my argument against three potential objections. §4 shows how to extend this argument to other theories of phenomenal concepts, using indexical and direct reference theories as examples. §5 outlines how these negative arguments yield a new and attractive form of physicalism, roughly a type-A phenomenal concept strategy.

§2 Constitutional Theories and the Phenomenal Concept Strategy

2.1 Constitutional Theories:

The Phenomenal Concept Strategy is an argument schema that must be supplemented with a theory of phenomenal concepts that vindicates the proposed conceptual dualism. It will be helpful to start with a concrete version. Once the structure of the argument is clear, we can see how it will generalize.

There are different ways you can think about phenomenal experiences like pain. You can think about pain in an indirect way – as the result of some neural state or as the experience you have when you stub your toe. But you can also think about it directly, not in terms of its causes or effects, but in terms of what it *feels like*. Imagine you stub your toe. While jumping about cursing, you think, "I hate *this experience*". In thinking this, the pain you're experiencing doesn't seem to stand at a distance from your thought. Rather, you use the pain in order to think about what you hate. It's not a contingent feature that you use to pick out the referent, but an essential one: the hurtiness of the pain.

These are the sorts of motivations that underlie the constitutional theory of phenomenal concepts.[5] On this view, phenomenal experiences aren't merely the referents of our phenomenal concepts; they're constituents, providing the concepts' cognitive significance and fixing the concepts' reference. On one way of spelling this out, phenomenal concepts have a quotational structure "the experience __" where the blank is filled in by a token of the

[5] See: David Papineau (2002), David Chalmers (2003), Katalin Balog (2012).

relevant type of experience. Katalin Balog suggests we think of this on the model of quotation marks: Much as ‘ “dog” ’ refers to the word of the type between the quotation marks, so the phenomenal concept will refer to the experience of the type presented between the ‘*’s of the “the experience” operator. Because phenomenal experiences are constituents of phenomenal concepts, phenomenal concepts carry with them the phenomenology of the experience. Knowledge involving these concepts therefore is supposed to give us unmediated insight into the “essence” of the phenomenal experience. (Balog 2012a, 9).

How does this help the phenomenal concept strategy? Take a familiar necessary a posteriori truth: Water is H_2O. Given a complete description of the microphysical facts, I’m able to determine (a priori) that water is H_2O, precisely because I am able to work out that H_2O plays the role a priori associated with ‘water’. (Given knowledge of the microphysical truths, I can work out that gloms of H_2O molecules will behave in such a way as to be liquid, clear, boil at 100°C, etc. and hence to be the watery stuff around here.) But if phenomenal concepts don’t refer via a priori associated descriptions – but rather via instances of the appropriate type of experience – there is alleged to be no possibility for a priori derivations of phenomenal-physical truths. This generates the epistemic gap. And because phenomenal concepts carry their phenomenology with them (whereas physical concepts don’t) physical concepts seem to leave out the “what it’s likes”. This is allegedly what makes an *explanation* of such truths seems impossible, generating the explanatory gap. While I take the constitutional theory to offer a highly compelling account of phenomenal concepts, which can give us much of what we want out of such a theory, I will argue that it cannot help type-B physicalism.

2.2 A Priori Derivation Without Associated Descriptions:

I want to put pressure on the idea that a priori derivations require that one of the terms involved employs an a priori associated description. I’ll offer an alternative picture that plausibly facilitates an a priori derivation, even on the constitutional model. The basic idea is this: Just because phenomenal concepts don’t refer via associated descriptions doesn’t mean that there’s nothing to use in an a priori derivation. The concepts involve instances of the phenomenal experiences themselves. If these experiences simply are physical processes, then

it should be possible for the physical system they're processes of to be constructed in such a way that the processes can be reliably matched up to information stored *describing* the processes. And that would yield an a priori derivation of the phenomenal from the physical.

According to the phenomenal concept strategy, there's an ideal epistemic gap between the physical truths (P) and the phenomenal truths (Q). That is, it's *ideally conceivable* that there be a world physically just like our own, but phenomenally completely different (that P & ~Q). What does it take for some proposition Z to be ideally conceivable? One way of getting at ideal conceivability is by way of what an *ideally rational* agent could conceive of. Since this raises the further tricky question of what it takes to be an ideally rational agent, I'll stick with a more minimal characterization: Z is *ideally conceivable* iff it's not knowable a priori that ~Z.[6] How should we understand the notion of a priority employed here? Chalmers characterizes a priority as follows: "If I cannot know that P independent of experience, but another less limited being could do so, then it is a priori that P." (Chalmers 2002) What I take to be relevant for Z being a priori is that *some creature* could come to know, independently of experience that Z. It follows that Z is ideally conceivable iff it's not the case that some creature could know that ~Z independently of experience.

Type-B physicalists claim that it's ideally conceivable that P & ~Q: that it's not the case that some creature could know that ~(P & ~Q) independently of experience. I will show that contrary to their claims (if physicalism is true), some creature could know that ~(P & ~Q) independently of experience.

The structure of this section is as follows: In order to assess PCS (and ultimately type-B physicalism on the whole) on grounds that its defenders will accept, I will assume the truth of physicalism. I will design, in low-level, sub-personal terms, an agent who (1) has complete general physical knowledge, (2) has concepts that meet the constitutional theory's requirements for being phenomenal concepts, and yet (3) is wired such that she could immediately determine the phenomenal-physical truths.

If I can construct such an agent, it will show that the constitutional theory of phenomenal concepts does not entail an ideal epistemic gap as maintained by type-B physicalism. A less limited agent with the requisite concepts would not find a gap. It follows

[6] This is roughly Chalmers's (2002) definition of ideal negative conceivability.

that we find these gaps due to contingent features of our own psychology.

2.3 An Alternative Psychology:

We entertain all sorts of thoughts, form judgments on all sorts of matters, and sometimes come to know things a priori. In all these cases, there is low-level processing occurring in our brains, in virtue of which we think and know what we do. When we consider an agent, we can describe these thoughts in higher-level terms (which might involve phenomenal concepts) or in terms of the lower-level, subpersonal workings of their brain (which will always involve physical concepts). If physicalism is right, both descriptions capture the same thing.

We should be able to give a complete description of what happens, at a low level, in Mary's brain when she successfully does arithmetic. Likewise, we should be able to give a complete low-level description of what happens in her brain when she thinks about phenomenal redness (using phenomenal concepts) or when she reasons using these concepts. Type-B physicalists, qua physicalists, must grant that such a low-level description of Mary's brain in physical terms would be *complete* (even if it could be alternatively described using different concepts). So if I can give a description of how someone in Mary's position could come to know the phenomenal-physical truths a priori, *even if this story is told in low-level physical terms*, it would follow that this person had bridged the epistemic gap.[7]

This is precisely what I propose to do. I'll start by describing a computer analogue of Mary.[8] I'll argue that – if wired correctly – such a creature could determine, in a principled way, that its "phenomenal concepts" and "physical concepts" corefer. Having described this coreference mechanism in low-level terms, I'll imagine a conscious agent with a psychology like that of the computer. In this agent, the states embedded in the "phenomenal concepts" are phenomenal states; the low-level coreference mechanism results in conscious judgments

[7] The claim is that – if Mary functioned in this way – *she* would bridge the epistemic gap. It does not follow that our appreciation of her will enable *us* to bridge the gap. But this is sufficient for establishing the falsity of type-B physicalism.

[8] This structure was inspired by Perry's (2001) proposal for the psychology of indexical knowledge. I discuss Perry's version of the PCS in §4.1 and show that the same problems apply to it.

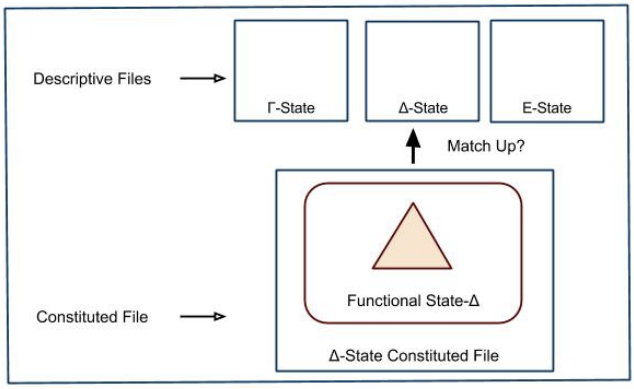
of the form "*This experience* [phenomenal concept] just is this physical state [physical concept]". In the sections that follow, I'll argue that these judgments amounts to a priori knowledge. In this section, I'll simply (i) highlight the features of the psychology that will prove crucial for defending the claim of a priori knowledge, and (ii) explore what would follow if this creature could indeed determine the phenomenal-physical truths a priori.

Let's begin with a computer analogue of Mary. Imagine concepts as files that store information about their referents. To be an analogue of Mary, the computer must have complete general physical knowledge, and so it has files describing every state the computer could be in.[9] Call these files *descriptive files*. These files are analogous to, e.g., captive Mary's (physical) concept of phenomenal redness, which – by hypothesis – picks out phenomenal redness in terms of the physical state of agents experiencing it.

The computer also has analogues of phenomenal concepts, which I'll call *constituted files*. Just as my phenomenal concepts contain tokens of the relevant types of phenomenal experiences (which, if physicalism is true, just are certain physical states), the computer's constituted files will contain tokens of the relevant types of computer states.

To have these "concepts", the computer must have an analogue of the "the experience" operator, which enables its own internal states to operate as the contents of files. Imagine that the computer is in internal state Δ, and that the Δ serves as the contents of a file, creating a constituted file of the Δ. As I've designed it, there is also a descriptive file that stores a complete description of Δ.

[9] This isn't quite right. If the computer had files describing *every* state it could be in, we'd get explosion. It would need to have files describing not only the state-A, but also the state of describing the state-A, and the state of describing the state of describing the state-A, and so on. But all we really need for this argument to work is that the computer contains descriptions of the first-order states that are equivalent to phenomenal states in conscious creatures. It needn't store descriptions of what it's doing when it recognizes a conscious state, recognizes that it's recognizing a conscious state, etc.



We now have a computer analogue of the mind of captive Mary when she is presented with a red piece of paper for the first time. Both Mary and the computer have complete general physical knowledge.[10] Mary has a physical concept of phenomenal redness (a concept *describing* the state that subjects are in when having red experiences). Our computer has a descriptive file of Δ (a file *describing* Δ). Mary has a phenomenal concept of redness (a concept that is *constituted by* her experience of phenomenal redness – which, by hypothesis, is a physical state). The computer has a constituted file of Δ (a file that is *constituted by* an instance of Δ).

If the structure of phenomenal concepts is sufficient to generate an epistemic gap between the phenomenal and the physical, then the computer will – like Mary – be unable to match up Δ (in the constituted file) with the description of Δ. Must this be the case? I will argue that it need not be: The computer could be constructed such that there's a reliable low-level mechanism by which the constituted file is matched to the corresponding descriptive file because the two files "corefer".

How could this be? At first pass, one might think it's obvious that a computer could be set up to match the constituted file to the descriptive file: It could simply be programmed to recognize and merge the two files when they were present. But this will not help us. This is analogous to insisting that we might have an innate disposition to match up our phenomenal experiences to certain internal states: an innate disposition to recognize our phenomenal experiences as instances of a certain physical type. While we could have such an innate disposition, if there is no principled connection between the concepts we are matching up –

[10] More carefully, the computer stores complete descriptions of all the physical truths. I don't mean to imply that the computer has intentionality.

simply a brute tendency (programming) that causes us to do so – the resulting beliefs will not amount to a priori knowledge. We could just as well be "programmed" to mis-match phenomenal and physical concepts. The problem is that the match-up process is *arbitrary*. Consider a computer that is programmed to match shapes with specific English terms for those shapes (e.g. 'equilateral triangle', 'acute isosceles triangle', 'right isosceles triangle', 'rhomboid'). Suppose we modify a shape, transforming it from a right isosceles triangle to a right scalene triangle. The only way for such a computer to accurately match shapes to words over this transformation is for it to be programmed as a look-up tree, for there is nothing whatsoever that connects English shape descriptors to the shapes, apart from arbitrary convention.

We need to show that our brains could be wired to match up the phenomenal concepts and the physical in a *non-arbitrary* way, because they corefer. I think that this is possible. If I draw a triangle, there are several ways you might represent what I've drawn. You might represent it by drawing another triangle. A computer could doubtless match these drawings up, in a non-arbitrary way, responding to their natural similarity. However what we have here is two instances of a triangle, not a triangle and a description. The match-up is trivial. And, as we've seen, you could describe the triangle by writing the word 'triangle'. The computer could also match the drawing and the word. But because the connection between the two is completely arbitrary, the human analogue wouldn't qualify as a priori knowledge.

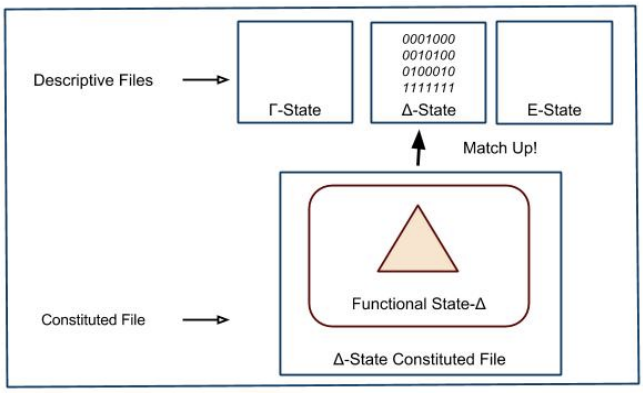
However, you could also describe the triangle by using propositions to describe a two-dimensional array of 0s and 1s, isomorphic to a triangle. A computer could doubtless be programmed to recognize this as a triangle. Here, the relationship between the triangle and the description is not trivial, as in the image case. But it is also not arbitrary, as in the 'triangle' case. Due to the analog relationship between the two files, the match-up process is robustly reliable across counterfactual worlds. (If we made a small change to the constituted file, there would only be a matching-up if there were a corresponding change in the descriptive file.) This counterfactual reliability is not due to the brute-force programming of a look-up tree. Rather, the computer programming is *making use of* a natural connection between shapes and descriptions of two-dimensional arrays to yield a match-up that is robustly counterfactually reliable. We thus have an example in which there's a principled

relationship between a description of a state and the state itself, based on which recognition could occur. (More on the relevant relationship and how it facilitates a priori knowledge will follow.)

Return to our computer analogy. Suppose that the information describing Δ (stored in the descriptive file) and Δ itself (stored in the constituted file) are formatted such that they stand in the sort of principled relationship just described. A computer could – if properly programmed – pick up on this relationship, recognizing the descriptive file as describing the very state contained in the constituted file, in a robustly reliably way.

One might worry here that I've said that the computer must be "properly programmed" to pick up on the relevant relationship. Doesn't "programming" mean arbitrariness? It doesn't. Programming is required to compute that 1+1=2, or any other thing the computer might process. Similarly, any conscious judgments I make are the result of low-level computations done by my brain, and are dependent on the details of how my brain functions – roughly, the "program" it uses to process information. What would be worrying is if (unlike the case I've described) the programming yielded judgments in a brute and arbitrary way, as a lookup tree might.

The figure below gives a simplified, but concrete illustration of the computer architecture I'm imagining. The computer has "descriptive files" that store descriptions of various states the computer could be in. One of these files describes the state Δ. There's also a "constituted file" of Δ, in which an instance of Δ is embedded. The formatting of these files is such that they stand in the sort of principled relationship described above. For simplicity, let's imagine this using the "triangle" model given above.



We might imagine Δ to be a state in which computer circuits are literally activated in a triangle pattern, and the descriptive file to be a file that describes a two-dimensional array activated analogously. Alternatively, we might imagine Δ to be a representation of a two dimensional array, activated like thus-and-so, and the descriptive file to be a file storing instructions for drawing a triangle (a program that, when followed, results in the creation in the neural circuitry of a representation of a certain type of two-dimensional array).[11] The goal is simply to have (i) the constituted file contain the relevant internal state, and (ii) the descriptive file describe that state in a way that has a principled but nontrivial connection (such as an isomorphic relationship, of the sort described above) to the internal state itself. While the state Δ is a toy example, simpler to wrap our heads around than actual phenomenal states are likely to be, it illustrates the basic principle.[12]

I claimed that we can construct a computer analogue of Mary, who could match up constituted files with their corresponding descriptive files, in such a way that a conscious creature with this psychological structure could know the phenomenal-physical truths a priori. As a first step towards this, I've sketched an example of a computer that can match up constituted files with their corresponding descriptive files in a way that does not merely rely on brute programming. I've claimed that the matching up of constituted and descriptive files in this computer is both non-trivial and non-arbitrary. Before moving on, I want to highlight the features of this match-up process that (i) account for its non-arbitrary nature, and (ii) will ground my further claim that a conscious creature with this sort of psychology could know the phenomenal-physical truths a priori.

Important features of the alternative psychology: (1) The match-up process I've described is robustly reliable across counterfactual worlds. If either the constituted file or the descriptive file had been structurally different, the match-up would not have occurred.

[11] There's no need for this program to be understood in intentional terms. There is no homunculus inside the computer *interpreting* the instructions. There is simply low-level code that "runs" the "create two-dimensional array like so" pattern. There's a sense in which the programming of these instructions might be thought to be arbitrary – the sense in which it's arbitrary that a computer program be written in C++ or Python – but *given that* the computer functions to "interpret" the programming in the way that it does, the connection between the "draw triangle" program and the two-dimensional array is not arbitrary.

[12] In fact, the idea of a state in which computer (or neural) circuits are literally activated in a triangle pattern isn't as ridiculous as one might think. The spatial layout of activity within our brains really is essential for processing and keeping track of information, and plays a significant role in the conscious experiences we have. (Groh 2014)

Suppose we make a small change to the constituted file. The constituted file will be matched to the descriptive file iff there's a corresponding change in the descriptive file, such that the match-up is reliable. (2) The match-up process is robustly reliable because of its unique internal structure. This is not like a case where a clairvoyant miraculously happens to reliably match up coreferring files – where Δ needn't factor into the explanation for why the match is made. In this case, the match-up mechanism literally contains and operates upon Δ, and this is integral in explaining why the computer forms this robustly reliable match-up. (3) There's an analog relationship[13] – a natural isomorphism – between the constituted and descriptive files. While this may not be crucial to the match-up yielding a priori knowledge, it helps to make vivid how the match-up process could be robustly reliable in a way that relies on the internal structure, and so to make the example I've sketched compelling.[14]

Thus far, I've been describing a computer. I've given low-level descriptions of the various internal states it can be in, and described a certain sort of processing that could go on relating these states. But our ultimate goal is to make a claim about consciousness, and the possibility that an agent could come to know a priori that their phenomenal concepts (the concepts that – on the constitutional theory – carry their phenomenology with them) corefer with the relevant physical concepts. So let's now stipulate that the creature with the cognitive architecture just described is conscious. *The state Δ is a phenomenal state.* What follows?

As we've noted, subconscious processing underlies all of my conscious judgments, from

---

[13] The notion of analog *representation* is a familiar one. If you represent your age using 40 candles, this is an analog representation: If your age changes, the number of candles must correspondingly change in a linear fashion. This relationship is symmetrical: If your age were a representation of the number of candles on a birthday cake, when we change the number of candles on the cake, your age would need to correspondingly change. This sort of symmetrical analog relationship holds between the constituted file of Δ and the descriptive file. I have simply called it an analog *relationship* because it is not important in my example that either file be taken as a *representation* of the other one.

[14] Physicalists of a certain stripe might worry that this match-up process relies on the physical *structur*e of experiences. A creature wired up as described might judge that her pains were identical to *this* internal state (CFF), but she is doing so based on the internal structure or functioning of CFF, not the particular physical composition. The mechanism would have operated in precisely the same way, were we to have surgically replaced the C-fibers with identically functioning D-fibers. Does this mean that my challenge only applies to functionalists? I don't think so. Note that the ideal agent who is relevant to the debate between type-B and type-A physicalists will possess complete general physical knowledge. This will include knowledge of what neural (or silicon-computational) states realize particular functional states in humans, bats, iPhones, etc. Holding this knowledge fixed, someone with the psychology I've described will be able to (just by thinking) combine her a priori knowledge of phenomenal-functional truths with knowledge of the neural states that realize these functional states, thereby yielding knowledge of phenomenal-neural truths.

mathematical to moral. While we've given a subpersonal explanation of the processing underlying the creature's judgment, the result is a conscious judgment, employing phenomenal concepts: "*This experience* [phenomenal concept] just is this physical state [physical concept]". In the following subsection, I will argue that this judgment amounts to a priori knowledge. For now, let's take this for granted and see what follows.

The agent I've described requires a very particular sort of neural architecture – both in the formatting of their physical and phenomenal concepts and in the programming to match them up. Given the peculiarity and complexity of this cognitive architecture, together the lack of any apparent evolutionary benefit to being designed like this, our default assumption should not be to imagine ourselves "ideal" in this way. (We certainly don't seem to be.) If this is right, the phenomenal-physical identities will always seem arbitrary to us. But it is *a contingent feature of our psychology* – a quirk, perhaps a rational defect – that we can't directly determine the phenomenal-physical truths.

We still have an explanation of why *we* are unable to determine the phenomenal-physical truths a priori. But the explanation has to do with *us*, not with the a posteriori status of phenomenal-physical truths.

For a claim to be knowable a priori it has to be the case that some agent could come to know the truth of the claim independently of experience. And if what I have said is right, we have a sketch of just such a creature: a conscious creature who could immediately determine the phenomenal-physical identities (P=Q), much as we can immediately determine that 1+1=2. It follows that there is no *ideal* epistemic gap between the physical (P) and the phenomenal (Q): There's a possible agent that could work out a priori that ~(P & ~Q). But recall that the goal of type-B physicalism is to accept the ideal conceivability of P &~Q, while blocking the alleged metaphysical implications. If what I've argued is correct, the constitutional theory of phenomenal concepts doesn't provide a way to do this; rather, it provides an explanation of why we believe there to be such a gap, though there really is not one. While this may offer a plausible physicalist reply to anti-physicalist arguments, it does not support type-B physicalism. Essentially, we have arrived at a psychological story supporting type-A physicalism.

2.4 A Priori Knowledge:

Thus far, I have argued that there could be a creature whose psychology enabled her to immediately and directly "match up" any experience she'd had with the description of that experience. I've claimed that the match-up judgments formed by a conscious agent with such a psychology would constitute a priori knowledge, and have explored what the implications are if this is true. Now I want to step back and defend my claim that these immediate phenomenal-physical judgments qualify as a priori knowledge. (Recall that the type-B physicalist can happily accept that an agent with a radically different psychology could work out the phenomenal-physical truths a posteriori – even *we* can work out these truths a posteriori. What they deny, and what I must prove in order to refute their argument, is that a creature could determine the phenomenal-physical truths *a priori*.) I'll begin by arguing that the resulting belief is knowledge. Next I'll argue that this knowledge is a priori, rather than introspective a posteriori knowledge.

2.5 Is it Knowledge?

To simplify the discussion, let's imagine a particular case of a conscious agent with the psychology I've described. Imagine a subject who – like Mary – knows all the general physical truths, but has never had an experience of phenomenal redness. This agent differs from Mary simply in that she has a psychological structure analogous to the computer I have described. Call her Mary*. When Mary* sees red for the first time, she not only forms a new concept (of what this new experience is like), she also forms an immediate judgment that what this phenomenal concept picks out is the same kind of thing her physical concept picked out. The question we're now faced with is: Does this judgment – that thus-and-so experience is of thus-and-so physical kind – amount to knowledge?

Let's start by dismissing one reason you might think it wasn't knowledge. I have argued that a conscious creature, with phenomenal concepts like those described by the constitutional view, could have a physical constitution that enabled it to recognize immediately (without investigation of the world) that the phenomenal experiences it was having were identical to certain physical processes. But this recognition stemmed from low-level subconscious matching of the current states of the creature with information describing

physical states. I have not argued that the process by which the phenomenal-physical judgments are formed would be transparent to consciousness.[15]

One might object that the process by which we come to have knowledge must, at least in principle, consciously accessible. If this were the case, what I have described would not be a case where a creature *knows* that the phenomenal experience they're having is a certain physical state, but merely one where they have an immediate intuition to that effect.

But it's not clear that we do always have conscious access to the processes underlying our knowledge. Some immediate intuitions seem to constitute non-inferential knowledge: I seem to *know* that modus ponens is a valid argument form; I seem to *know* that pain is bad. How I arrive at these judgments is not something accessible to my consciousness. If these are to count as knowledge, then the process by which we come to have knowledge needn't be consciously accessible.

Why should the physicalist think that a creature like Mary* would *know* the phenomenal-physical truths? First, the physicalist is committed to holding that Mary*'s phenomenal-physical beliefs are *true*. She matches up low-level descriptions of her internal states with phenomenal experiences that – according to physicalism – are those very internal states. Furthermore, this Mary* doesn't just randomly happen to form true phenomenal-physical beliefs. She forms these beliefs *because they are true*. The low-level matching process doesn't arbitrarily match up the descriptive file with the constituted file – it matches them up *because* the state that fills the constituted file is the very type of state described by the descriptive file.

One might object that phenomenology is essentially conscious. If the match-up process is subconscious and cannot be consciously reasoned through, you might worry that Mary* isn't really judging *on the basis of the phenomenology* that there's a connection between the constituted concept and the descriptive concept. And surely the phenomenology of (e.g.) pain must play a role in coming to know that pain is thus-and-so

[15] Consider a skilled rower. An oar in water that would look bent to a layperson *looks straight* to the rower. But the processing that causes the rower to see the oar as straight is not consciously accessible to the rower. All that the rower is conscious of is the result of the low-level processing: the oar looks straight. Similarly, we might expect creatures with the psychology I've described to simply be aware of the result of the match-up process: a direct awareness that the experience they're now having is described by a certain physical concept.

physical state.

But I don't think the physicalist can justify this move. According to the physicalist, pain is essentially a certain physical state. Thus, when Mary* judges that a certain phenomenal-physical identity is true, she does so *based on the essential nature* of the phenomenal experience. There are not (assuming physicalism) two properties – the phenomenal and the physical – one of which Mary* is accessing, and the other of which she is blind to. There is only one property. And it is precisely this property which Mary* is using to deduce the phenomenal-physical truths. The physicalist cannot deny that judgments formed on this basis amount to knowledge. The only real question is whether this knowledge is knowledge that connects the phenomenal way of thinking about experiences with the physical way of thinking about them. But this question concerns the *conscious outputs of the judgment*, not the mechanism by which the judgment is formed. And the conscious output of Mary*'s low-level matching-up is a judgment that connects the phenomenal way of thinking about experiences (using phenomenal concepts) to the physical way of thinking about them.

Finally, while I have argued that the match-up process is robustly reliable, my argument is not grounded in a general assumption of reliabilism. What's epistemically important about the match-up case is not the mere reliability of the mechanism of belief, but the fact that the agent is determining phenomenal-physical identities based on internal access to the entities in question. The mechanism by which Mary* forms her phenomenal-physical judgments is reliable precisely *because* it is partially composed of the judgment's truth-makers: The mechanism doesn't merely track some external fact that Q = P. Rather, the mechanism literally contains the phenomenal/physical state (Q/P), where this state itself is integral in explaining why Mary* reliably judges that Q = P.

It might be illuminating to compare the match-up case to a case involving mathematical intuition. My aim is not to make any general points about justification or knowledge, but merely to make it plausible that the particular psychology I've described can yield knowledge, by highlighting the similarities between the match-up case and a case of justified mathematical intuition.

First consider a mathematical clairvoyant, who has mathematical truths reliably but inexplicably pop into her head. Her beliefs presumably are not justified and do not constitute

knowledge.[16] Now consider (justified) mathematical intuition. The mathematician performs a subconscious mental operation, using a mental model that mirrors the structure of the natural numbers. There must be *something* epistemologically important about the process by which the mathematician's intuitions are formed. What's important is not simply the reliability of the intuition, as that's something that the clairvoyant shared. Arguably, what's importantly different about this case is that the mathematician's intuitions are reliable *because* the structure of the internal process mirrors the structure of the mathematical truths. In this case, as with Mary*, what's relevant is that the process by which the judgment is formed essentially makes use of the truth-makers (or, as in the mathematical case, models that are isomorphic to them).

So we can happily grant, contrary to a simple reliabilist, that a phenomenal-physical clairvoyant wouldn't have phenomenal-physical knowledge. Moreover, my arguments don't imply (as a general commitment to reliabilism would) that it's possible for an agent to differ in epistemic status from her duplicate. In the case at hand, any intrinsic duplicate of Mary* will also be an epistemic duplicate regarding her knowledge of the phenomenal-physical truths.[17]

2.6 Is the Knowledge A Priori?

A further challenge lies in showing that the knowledge is a priori. Type-B physicalists have no difficulty accepting that some creatures can come to know phenomenal-physical truths. (They take us to be such creatures.) What they cannot accept is that such truths can be known *a priori*. And it seems plausible that, contrary to my claims, what I've described is really a kind of introspective a posteriori knowledge.

---

[16] If you think that the mathematical clairvoyant *does* know the mathematical truths, you should straightforwardly be on board with Mary* knowing the phenomenal-physical truths.

[17] While my argument is compatible with a weak internalism of the sort that simply requires that intrinsic duplicates be epistemic duplicates, there are stronger versions of internalism that do not sit well with my argument. In particular, I have not argued that Mary* will be able to consciously distinguish her own position from that of a clairvoyant who simply has the phenomenal-physical truths pop into her head. If justification required that one have access to the belief-formation process, Mary*'s beliefs would not be justified. (Likewise, I presumably would not be justified in my belief that modus ponens is a valid argument form, as I do not have access to the process by which I came to form this belief.) But even if this provides a way to avoid my conclusions, it comes at the cost of forcing the type-B physicalist to adopt a very particular story about epistemic justification, which many will not find attractive.

Consider Descartes's cogito. "I think, therefore I am" can be known through introspection alone, but nevertheless is plausibly only knowable a posteriori.[18] Similarly, my knowledge that *I'm currently thinking about a priori knowledge* is something I know through introspection alone, but also seems a posteriori. How can I hold that Mary*'s knowledge of the phenomenal-physical truths is a priori if I accept that these other cases of introspective knowledge aren't? What is the relevant difference between these two cases?

In cases of introspective knowledge that we intuitively feel to be a posteriori, the knowledge is based on contingent truths that are known through introspection. Descartes's cogito takes *as a premise* the contingent truth "I think"; it's his observance (through introspection) of this contingent fact that grounds his knowledge. When I think that *I'm currently thinking about a priori knowledge*, my knowledge is grounded in the contingent fact that I'm currently thinking about a priori knowledge. The contingent fact that I'm currently thinking this *justifies* my belief. But Mary*'s subconscious determination of the phenomenal-physical truths is very different. Though it's true that in order to make such judgments, Mary* must have had certain phenomenal experiences (which she's only contingently had), these experiences play a very different role in her beliefs. Mary* must have had a red experience in order to have the relevant concepts to think about the phenomenal-physical truths. But Mary*'s judgment isn't grounded in the contingent fact *that she's had a phenomenal red experience*. Her judgment is rather grounded in the *necessary* (assuming physicalism is true) fact that such-and-such phenomenal state is the physical state that it is. Mary*'s having had the relevant experience plays an enabling role for her to be able to work out the phenomenal-physical truths, but – unlike clear cases of a posteriori introspective knowledge – it does not play a justificatory role.

There's a second related argument that might be made against the a priori status of Mary*'s knowledge. One might argue that the "match up" process that I've described involves a kind of *inspection* of the agent's concepts, and that knowledge come to through inspection (whether inspection of something external to the agent or something internal) always yields a posteriori knowledge. On this line of thought, there's an important difference between a priori conceptual analysis (which involves *employing* concepts) and a posteriori

[18] One who takes the cogito to be knowable a priori will doubtless happily grant that Mary*'s knowledge is also a priori.

inspection of concepts.

It's hard to get a grasp on the difference between knowledge come to through inspecting concepts and knowledge come to through employing them. (When I conclude e.g. that a bachelor is a marriage-eligible man, does this involve *inspection* of my concept or *employment* of it? How can we determine the difference?) However, there is one clear difference between inspecting a concept and employing it: When I inspect my concept of redness, I can thereby come to know that *I have a red concept*. By contrast, I can't conclude this simply by employing my red concept. It is clear that one can only know that they have a red concept a posteriori. This might seem to be reason to think that there is a principled difference between inspection and employment of concepts, and that inspection only yields a posteriori knowledge.

While I agree that one can only know whether they have a red concept a posteriori, I think that this can be explained in the very same way as the cases of introspective a posteriori knowledge discussed above. My knowledge that I have a red concept is a posteriori because it is grounded in the contingent fact that I have a red concept. But once again, the phenomenal-physical knowledge that I've described doesn't fit this model. Mary*'s knowledge of phenomenal-physical truths *presupposes* that she has the concepts necessary for thinking the relevant thoughts. (It's a precondition for employing a concept that you possess it.) But her knowledge isn't grounded in the presupposition that she has the said concept. Rather, it's grounded in the necessary truth that the relevant phenomenal and physical concepts corefer. To put this another way, if we were to flesh out the reasoning that underlies Mary*'s coming to believe that phenomenal redness = thus-and-so physical state, *the fact that Mary* has a red concept* would not be a premise. The source of her judgment's justification lies entirely in the noncontingent fact that the phenomenal state embedded in the phenomenal concept just is the state described by the physical concept.

Since it's so difficult to get a grasp on the difference between employing and inspecting concepts, we might characterize the difference by relying on the clear-cut cases where there is a difference between the two processes. If we follow this suggestion, it's natural to think that the marker of judgments made by *inspecting* concepts is that the contingent fact of the concept's existence plays a role in the judgments. But if this is how we analyze the

inspection/employment distinction, Mary*'s judgment would not involve inspection.

So we haven't seen a reason for thinking that Mary*'s knowledge is a posteriori. It's possible that one could (1) give a radically different account of the difference between employing and inspecting concepts, on which inspecting concepts always yielded a posteriori knowledge, and (2) show that the knowledge I've described involves concept inspection on this account. But it's difficult to see how such an argument could go.[19]

§3 Objections

I'll now move on to consider three objections to the argument I've sketched. (1) I consider an argument that the agent I've described is employing different concepts from ours, and hence says nothing about what's conceivable using our concepts. (2) I consider an objection that – if arguments against PCS are to take the view seriously – they cannot rely on "mere architectural" accounts of phenomenal concepts. One might think this will cause trouble for my argument, as it is merely the architecture of phenomenal concepts that I use to argue that Mary* can come to know the phenomenal-physical truths a priori. (3) I consider whether the agent I've described would find an *explanatory* gap.

3.1 Constitutional Files Aren't Analogous to Phenomenal Concepts:

Consider Mary and Mary*. When Mary is shown a red piece of paper for the first time, and forms a new concept of what it's like to see this color, she can't work out a priori that the experience she's now having is an experience that her physical red concept applies to. When Mary* is shown the red piece of paper, she immediately recognizes the experience she's

[19] Further, there is some intuitive reason to think that such an account could not succeed. Suppose I'm faced with a red block, an orange block, and a blue block. Rather than attending to the blocks, I decide to reflect on the *phenomenal experiences* that I'm having as I look at them. I think to myself: "That experience (red) is more similar to that experience (orange) than to that experience (blue)." If Mary*'s "match up" of the phenomenal and the physical counts as *inspection*, then it's hard to see how this scenario could fail to count as inspection: Surely what I'm doing when I make this judgment is detecting a commonality between my red and orange experiences, much as Mary* detected a commonality between the physical and phenomenal concepts. Further, intuitively what I'm doing is *inspecting* my experiences to determine how similar they are. But intuitively, once we hold fixed that I'm having red, blue, and orange experiences, my further knowledge that red experiences are more like orange ones than like blue ones is *a priori*. So if this involves inspection of experiences, then not all things we know through inspection are a posteriori.

having as a red experience. Call the phenomenal concept Mary forms RED, and call the phenomenal concept Mary* forms RED*. One might now object that – though RED and RED* refer to the same experience and have the same cognitive significance insofar as both contain an instance of phenomenal redness – RED and RED* are different concepts. RED* facilitates different a priori derivations from RED, showing that something about their cognitive significance differs: While it can be determined a priori that RED* and the physical concept of phenomenal redness $RED_{(PHYS)}$ co-refer, it can't be determined a priori that RED and $RED_{(PHYS)}$ co-refer.

For the sake of argument, let's grant that RED and RED* are different concepts. Would this help the phenomenal concept strategy, and hence type-B physicalism? RED* is clearly a phenomenal concept: It has the very structure proponents of the constitutional view put forward as the marker of phenomenal concepts. It not only refers to phenomenal experiences, but also characterizes them in terms of what it's like to have them. It carries with it the very phenomenology that it refers to. In fact – assuming that the state it "quotes" is a phenomenal state – these concepts seem to give us everything we could want from a theory of phenomenal concepts.[20] If the coreference of $RED_{(PHYS)}$ and RED* can be known a priori, then it is not the case that there's something special about phenomenal concepts that prevents the phenomenal-physical truths from being known a priori. Rather there is something special about a particular class of phenomenal concepts. What class? The only way to describe it is to say: "There is a class of phenomenal concepts that is unique insofar as these concepts can't – simply because of the psychology of the organism they're possessed by – be a priori determined to co-refer with physical concepts."

While I have said that Mary can't determine the phenomenal-physical truths a priori because a quirk (irrationality) in human psychology prevents us from matching up the phenomenal concepts with the physical, the objector will say: Mary can't determine the phenomenal-physical truths a priori because her phenomenal concept RED is a special kind of

[20] Balog (2012a) offers a list of desiderata that any adequate account of phenomenal concepts will satisfy. Apart from the desiderata specifying that phenomenal concepts should generate an epistemic/explanatory gap, red* meets every criterion. Relying on Balog's desiderata, the only way one could deny that red* is a phenomenal concept is by insisting that phenomenal concepts *just by definition* generate an epistemic/explanatory gap – an extremely unappealing route to take. While the defender of PCS could add additional criterion, Balog's list already looks quite thorough. It's difficult to see what independently motivated criterion one could add.

phenomenal concept – a phenomenal concept embedded within a psychology that can't match it up with physical concepts. But these are simply two ways of describing the same thing.

What of the conceivability argument? Type-B physicalism aims to embrace the first premise of the conceivability argument while denying its conclusion: holding that zombies, while not possible, are ideally conceivable. I've argued that the constitutional theory of phenomenal concepts cannot help the type-B physicalist achieve this aim: If our ability to conceive of zombies results from a quirk of our psychology, and a creature who lacked this psychological quirk would be able to determine these truths a priori, then zombies are not *ideally* conceivable. But one might argue that on the present line of argument, creatures with this alternative psychology are simply using different concepts. Using these alternative phenomenal concepts, zombies cannot be conceived of. Using human phenomenal concepts, zombies can be (ideally) conceived of.

This objection highlights the question of what it takes to be an ideal agent. Consider two different methods of locating the ideal agent: the first takes a human being, *holds fixed certain basic features of their psychology*, and improves them. Call this agent the *Human Ideal.* The second way of idealizing starts from scratch and locates the absolutely most rational structure for an agent's psychology. Call this agent the *Absolute Ideal*. The Human Ideal will be able to conceive of zombies; the Absolute Ideal will not be able to conceive of zombies.

Is ideal conceivability a matter of conceivability to the Human Ideal or to the Absolute Ideal? You can *say* whatever you like here, but if you care about ideal conceivability because of its potential to tell you something about *the world*, it's clear that it's the Absolute Ideal that matters. When the Human Ideal and the Absolute Ideal diverge in what they find conceivable, the Human Ideal doesn't tell us anything substantial about *the world* beyond what the Absolute Ideal tells us. Rather, it tells us something about the peculiar structure of the human mind. Ideal conceivability is only of philosophical interest insofar as it tells us something about what's possible. But no one would think that the Human Ideal – when it diverges from the Absolute Ideal – would have any implications for metaphysical possibility. Hence Absolute Ideal conceivability is clearly the notion that dualists have in mind in the

conceivability argument. So acceptance of its first premise requires accepting that an Absolute Ideal agent could conceive of zombies.[21]

Regardless of whether you choose to describe the concepts I've sketched as "the same" as our phenomenal concepts or not, the upshot is the same: There is nothing special about phenomenal concepts that prevents phenomenal-physical truths from being known a priori. If physicalism is true, the explanation of this failure must lie in our own psychology, not in anything distinctive about concepts that carry the phenomenology of their referent with them. In the most powerful sense of "ideal conceivability" – the sense that dualists use, and the only sense of philosophical importance – zombies are not ideally conceivable.

3.2 A Merely Architectural Account:

One might worry that if we are to take the type-B physicalist's core commitments seriously, we cannot rely on a merely "architectural" account of phenomenal concepts. Balog makes just this point in her response to Chalmers's (2007) "master argument" against the phenomenal concept strategy:

> The physicalist explanation of the substantial grasp [of a phenomenal property via a phenomenal concept] crucially involves the fact that there is something it is like to have an instance of [the phenomenal property]. This means that the constitutional account couldn't be cast in physical or quasi-phenomenal terms and still explain our epistemic situation. … Neither a neurophysiological, nor a mere "architectural" description of phenomenal concepts – e.g. that they are constituted by instances of the referent – can explain the substantial manner in which we refer the phenomenal properties. (Balog, 2012b: 15)

But isn't what I have done simply to give a "mere 'architectural' description of phenomenal concepts"? Haven't I neglected to take seriously the *phenomenal* way of conceptualizing phenomenal concepts by arguing that they can be matched up to physical concepts based on *structural* features? And if so, am I not failing in my ambition to provide an argument against type-B physicalism on grounds that its proponents should accept?

Whether it is problematic to use a merely architectural description of phenomenal

[21] It's also obvious that the Absolute Ideal conceivability of zombies is what type-B physicalists intend to accept. If they merely wanted to grant the claim that a Human Ideal agent could conceive of zombies, there would be a simple route to denying metaphysical conclusions: Human Ideal conceivability, grounded in our fickle human psychology, clearly tells us nothing about metaphysical possibility.

concepts depends on what work is being done with the description. There are two distinct questions we might ask: (1) If I think about phenomenal concepts using a merely architectural description of them, will I find an epistemic gap? (Or if Billy thinks about phenomenal concepts in architectural terms, will Billy find an epistemic gap?) (2) If I think about Billy the Martian's phenomenal concepts using a merely architectural description of them, will Billy the Martian find an epistemic gap?

In answer to the first question, the type-B physicalist may reply that we will not find an epistemic gap. One only finds such a gap when employing concepts that present phenomenal experiences under a *phenomenal guise*: Whether *I* find a gap depends on how *I* am thinking about phenomenology. But this is completely compatible with thinking that the answer to the second question is 'yes'. How *I* conceptualize Billy the Martian's phenomenal concepts – whether I give a low-level characterization of them or a phenomenal characterization of them – is irrelevant to the question of whether *Billy* will find an epistemic gap.

What I have tried to do in this paper is to build up an agent from the inside-out, constructing their psychology "from the ground-up". I have not asked "Do we find a gap?" but rather "Would an agent that functioned in this way at the low-level find a gap?" While I've given a low-level, subpersonal characterization of the process through which Mary*'s phenomenal-physical judgments are formed, when Mary* reflects on what she knows, she doesn't do so by thinking about subpersonal mechanisms or neural programming. The judgments Mary* has as a result this processing are judgements like "Phenomenal redness [phenomenal concept] is thus-and-so physical state." These phenomenal-physical judgements involve phenomenal concepts, complete (if you're a constitutional theorist, like Balog) with the rich phenomenology such concepts carry with them. Thus, if her judgements constitute a priori knowledge, then *Mary** is indeed bridging the epistemic gap.

The moral of the story is this: Type-B physicalists are not committed to holding that *my* thinking about a merely architectural characterization of phenomenal concepts will yield an epistemic gap, as thinking about such a structural characterization doesn't involve *using* phenomenal concepts. But the *creature we describe* using these architectural terms must find an epistemic gap. This is because, according to physicalism, an architectural characterization can give a complete characterization (in one set of terms) of a creature who possesses phenomenal concepts. Physicalists must grant that Mary* – however she may be

described – wouldn't find an epistemic gap. Type-B physicalism is untennable.

3.3 An Explanatory Gap Remains:

One might think that the creature I have designed would still find an *explanatory* gap. I haven't said anything that explains *why* the phenomenal-physical truths hold. Wouldn't Mary* still wonder *why* the phenomenal was identical to the physical? Wouldn't she still feel that these identities were somehow *arbitrary*? And wouldn't that show that there's still an ideal explanatory gap left open?[22]

First, it's important to get clear on the sense in which consciousness seems to leave open a problematic explanatory gap. There are many cases where (i) we know something to be true a priori, (ii) we cannot give any further *explanation* for this truth, and yet (iii) we don't find there to be a problematic explanatory gap. This happens whenever we take ourselves to have hit bedrock in our explanation. Consider the following cases: modus ponens is a valid form of argument; pain is bad; $a = a$. I seem able to see through to the truth of each of these statements directly. Yet I can't explain *why* any one of them is true. Despite this, I don't find myself up at night, wracking my brains for an answer to the question "Why does $a = a$?" This doesn't seem arbitrary or to call out for explanation. We've hit bedrock.

If physicalism is true, then the phenomenal-physical truths are another example of hitting bedrock. There is simply nothing more to be said (cf. Papineau 2011). This raises a further challenge: Suppose that Mary* begins to contemplate the possibility of dualism. She reasons "Yes, I immediately intuit that P=Q, but perhaps *that's* just do to some rational defect of mine matching up one brain-state to another. Perhaps phenomenal experiences are really non-physical!" Mary* would, in this case, be questioning whether P=Q really was a bedrock truth. One might ask whether in such a case, Mary* would be running up against an explanatory gap of the sort that the type-B physicalist maintains that even ideal reasoners would face.

But physicalists should think of this precisely on the model of $a = a$. First, given Mary*'s psychology, her questioning whether P=Q would be as bizarre and inwardly

[22] While I think this section gives a compelling response to the explanatory gap, this section of the paper is not essential to any of the previous arguments. If this paper has "merely" managed to give an explanation for why physicalists cannot accept an ideal epistemic gap, this strikes me as a good day's work.

unintelligible as my asking "is pain really bad" or "is *a* really identical to *a*"? Mary* would find the possibility of dualism only as intelligible as we would find the possibility that $a \neq a$. (We might imagine a Cartesian demon deceiving us into thinking this falsely, but can no more wrap our heads around the possibility than that.) Second, we can note that even if she were to entertain this possibility, there's no more a problematic explanatory gap here than there would be if someone began (bizarrely) questioning whether it was really true that $a = a$, and feeling that they needed an explanation of this putative fact. The important question is not whether someone could persuade themselves that there was a need for an explanation (where none seems readily found). Rather it's whether there *really is* a need for an explanation (where none can be found). The latter is problematic. But there's no reason to be bothered by the former. If physicalism is true, then even if Mary* begins asking *why* her phenomenal-physical judgments are true, this does not demonstrate that there's a problematic explanatory gap.

§4 Extending the Argument

Suppose that the constitutional theory is not the true theory of phenomenal concepts. Can we still construct an agent who can work out the phenomenal-physical truths a priori? In this section, I'll briefly introduce the indexical and direct reference theories of phenomenal concepts. I think these theories fail to offer what we want out of a theory of phenomenal concepts, and hence cannot support type-B physicalism for independent reasons. But I'll aim to show that even if they were the correct theories of phenomenal concepts, they cannot support type-B physicalism for much the same reasons the constitutional theory could not.

The basic idea is that any successful theory of phenomenal concepts will require agents to store information in their heads (e.g. that guides their ability to recognize the relevant sorts of experience, that guides their ability to imagine the experiences, that they can mentally point to and refer to). If the agent is just a physical system, and this information is stored in some physical state, then – by the same lights as the original argument in §2 – it should be possible for the physical system of the agent to match this state up with concepts describing the state.[23]

[23] The indexical, direct reference, and constiutional theories are not exhaustive of all the possible theories

4.1 Indexical Theories:

Indexical theories take phenomenal concepts to function as internal demonstratives. According to these theories, to form a phenomenal concept of the experience of redness is to internal demonstrative pointing to *that experience (whatever it is)*. Reference to the experience is rigid, picking out the same type of experience in each counterfactual world (just as reference to "that (pointing at a cup)" is rigid). But, contrary to the constitutional theory, phenomenal concepts are not individuated by the experiences they refer to. Had I *actually* been demonstrating a different experience, the concept would have referred to a different experience (just as "that" would refer to a bowl had I actually been pointing at a bowl).

Chalmers (2003) and others have argued compellingly against indexical theories, but here I want to consider whether, *assuming it offers a successful theory of phenomenal concepts*, it can do the work of supporting type-B physicalism. I'll argue that it cannot. I'll use Perry (2001) as my target, as he offers the most developed version of the theory.

Perry offers a thorough account of what is going on psychologically when we use indexicals. He proposes that our psychology has a three level structure. (Perry also uses an analogy to a computer, with concepts imagined as computer files.) The top level of our belief architecture is "detached" files, containing information gained from books, lectures, calendars, and the like. These files don't tell us anything about our current perceptual experiences and, by themselves, won't help us to navigate the world. The bottom level of the architecture contains "perceptual buffers" which temporarily store information gained from perception until it can be attached to a concept in the top level. When I see a cup and form an indexical concept *that* [cup], the experiences I'm having of the cup are stored in my perceptual buffer. Recognizing the cup *as a cup* involves connecting the perceptual buffer to a detached cup notion. To form such a connection, there must be a detected commonality between the information stored in the perceptual buffer and that stored in the detached notion.

---

of phenomenal concepts one could hold. Lycan, for instance, holds that phenomenal concepts are semantically primitive and essentially perspectival mental representations. But the extension of my argument to the indexical and direct reference theories should make clear how the basic strategy can be extended.

Similarly, when Mary has a red experience for the first time, the experience comes to fill her perceptual buffer. On Perry's view, this is what it takes for Mary to have a red phenomenal concept. This is not, by itself, sufficient for recognition of the experience as the same experience she'd read about in textbooks. To get that recognition, and the substantial knowledge that comes with it, Mary needs some basis for linking the red perceptual buffer to her detached red notion. If simply having a red experience in her perceptual buffer were sufficient for linking the buffer and the detached notion, then Mary could work out the phenomenal-physical identity a priori. Perry, rightly, thinks that Mary could not do this. He suggests that for recognition, Mary would need to e.g. (1) have a red detached notion that included the knowledge that tomatoes cause red visual experiences (2) see a red tomato, and (3) based on the recognition that a tomato was causing her experience, link up the perceptual buffer with the detached red notion.

But if my earlier arguments are correct, we can easily apply them to Perry's indexical psychology as well. Notice the direct parallels between the psychology Perry sketches and the psychology I sketched in §2. Both Perry's agent and Mary* contain "files" storing descriptive information about phenomenal experiences. On both pictures, there's a file that contains instances of phenomenal experiences themselves. (Perry calls this file the 'perceptual buffer'; I called it a 'constituted file'.) The only difference between the two pictures is in how this latter type of file/concept is individuated. On the constitutional picture, it's individuated by reference to the experience contained in the file; on Perry's it isn't. Because of this, the theories differ in what they take phenomenal concepts to refer to in counter-actual worlds. But this is merely a difference in philosophical interpretation, not a substantial difference in the agents' psychologies. Nothing on Perry's account differs from the constitutional account that could affect whether the agents could directly match up the files containing instances of the experiences with those describing them.

Perry thinks that match up between the red detached notion and the perceptual buffer requires, e.g., recognition that the perceptual buffer includes an experience of a tomato, together with knowledge that tomatoes are red. But if the argument from §2 is correct, we can design a creature for whom the similarity between the detached notion and the perceptual buffer itself is sufficient for such a match. While Perry is surely right that *Mary* could not

directly match up the experience in her perceptual buffer to her detached notion, this is again simply due to Mary's psychology. The indexical theory doesn't support the existence of an ideal gap, and so cannot support type-B physicalism.

4.2 Direct Reference Theories:

Direct reference theories take phenomenal concepts to be recognitional concepts that refer to phenomenal experiences directly, without employing any modes of presentation at all.[24] After seeing several cacti of a certain variety, I might form a recognitional concept "that sort of cactus" or "one of those". Similarly, according to direct reference theories, phenomenal concepts involve the ability to reliably identify experiences as "another one of those." But phenomenal concepts on this view are importantly different from standard recognitional concepts: My recognitional concept of cacti crucially involves a contingent mode of presentation (the way the cacti look) which facilitates my recognition of the cactus as "one of those" and gives the concept its cognitive significance. By contrast, phenomenal concepts on this view involve no modes of presentation. (It is not that they latch onto their referents using an *essential* mode of presentation – the way the phenomenal experience feels – whereas standard recognitional concepts use a contingent one. Phenomenal concepts on this view have *no modes of presentation whatsoever.*[25])

And though direct reference phenomenal concepts are sometimes described as demonstratives, they do not function as demonstratives standardly do: Unlike the indexical view, direct reference phenomenal concepts are individuated by their referents, and so refer to the same type of experience in every world considered as actual. As Levin puts it, phenomenal concepts "have *no* reference-fixing 'modes of presentation' *or* Kaplanian 'characters' that can change reference from world to world" (Levin 2007, 89). On the most simple direct reference theory, all it takes to have a phenomenal concept of redness is to have this sort of recognitional concept.

[24] Versions of the direct reference theory have been defended by Levin (2007) and Tye (2003).

[25] I'm not sure what sense can be made of these theories. How can I reliably discriminate without any mode of presentation whatsoever? Surely *something* must be guiding my recognitional abilities. Perhaps the way to make sense of direct reference theories is to take them as simply denying that the basis for my ability to discriminate is *consciously accessible.* We might think of this as a subconscious mode of presentation guiding my recognitional capacity.

Because phenomenal concepts have no modes of presentation, there is supposed to be no possibility of an a priori derivation of the phenomenal from the physical. According to this theory, phenomenal concepts effectively function as pointers directed at (and reliably caused by) a certain type of experience. These concepts are supposed to be so "thin" that there's no possibility of using them to determine phenomenal-physical truths; there's simply nothing there to lead us to these truths. This generates the epistemic gap.

But the direct reference theory cannot support type-B physicalism for reasons analogous to those I put forward against the constitutional theory. There must be some sort of information stored in my brain that guides my recognitional abilities and enables me to reliably discriminate experiences of a certain type. (My recognitional capacities aren't miracles, nor are they random.) While this information might not be consciously accessible, the argument I gave in §2 did not rely on the information used in the derivation being consciously accessible. By the same lights as the original argument, we could construct a creature whose brain stored information sufficient to recognize instances of a certain type of experience, who was "wired" such that this information could be immediately linked up to the (differently formatted) information describing phenomenal experiences in physical terms. We again find a gap between the phenomenal and the physical, in part, because of our psychology.[26]

§5 Conclusion

My main aim has been to show that type-B physicalism is untenable, on grounds that

[26] Tye (2003) formerly endorsed a more sophisticated version of the direct reference theory, which is similarly vulnerable to my argument. He argued that we should distinguish the question "What makes concept C a concept of e.g. phenomenal redness?" from the question "What makes concept C a *phenomenal* concept?" Just as on the simple direct reference theory, Tye thinks that concept C is a concept of phenomenal redness, just in case C is reliably triggered by (and because of) phenomenal redness. But this isn't sufficient to characterize what it takes to be a *phenomenal* concept. To be a phenomenal concept, C must be laid down in memory as the result of a phenomenal experience and must tend to trigger appropriate mental images in response to certain mental activities. But, as on the simple direct reference theory, in order for C to tend to trigger the appropriate sort of phenomenal experiences, the agent's brain must contain information guiding the construction of such experiences. Their brain must be able to automatically "unpack" this information, to generate an experience of the relevant type. But if their brain contains information that can guide the reconstruction of the relevant type of experience, then the agent's brain must store information such that some creature could (based on it) immediately determine these truths.

its proponents should accept. Whatever the correct account of phenomenal concepts is, we can develop an agent who has such concepts and is capable of matching them up to their physical concepts in such a way that the result is a priori knowledge. Since it offers an independently plausible theory of phenomenal concepts, I began by showing that such an agent could be constructed using the constitutional theory of phenomenal concepts. If the physicalist is right, phenomenal experiences are identical to physical states. I've argued that, if phenomenal concepts involve instances of the phenomenal experiences themselves, it should be possible for the physical system they're components of to be constructed in such a way that these states can be matched up to information stored *describing* the states.

I've described the psychology of such a creature, from the ground up, and have argued that the conscious result of such a low-level "match up" amounts to a priori knowledge that the descriptive and phenomenal concepts corefer: A creature with the psychology I've described could, given complete physical information, work out the phenomenal-physical truths a priori. It follows that there is not an *ideal epistemic gap*.

Finally, I've shown that my argument can be extended such that for any plausible theory of phenomenal concepts, we can construct an agent who doesn't find an epistemic gap. Once we acknowledge that possessing a phenomenal concept requires one's brain to store an instance of the relevant experience, or information sufficient to guide the creation of such an experience, or information sufficient to guide recognition of such experiences, we open up the possibility of a creature whose brain could directly match this information (or state) up to concepts describing the state. By the same lights as before, the result is a priori knowledge. Physicalists cannot accept that there's an ideal conceptual gap between the phenomenal and the physical. Physicalists cannot be type-B physicalists.

But while I've shown that phenomenal concepts cannot support the existence of an *ideal* epistemic gap, my argument is not at odds with what I take to be the true goal of the type-B physicalist: to offer a way of being a physicalist that respects our "dualist" intuitions. It just seems crazy to deny that captive Mary would be fooled if presented with a blue banana, to insist that if only we had more physical information, we would find zombies inconceivable, to insist that if only we knew more, we would cease to find phenomenal-physical identities arbitrary.

If what I have argued is correct, the physicalist must accept that an *ideal* agent wouldn't be fooled by the blue banana or find zombies conceivable. But, plausibly, our intuitions were not shaped by consideration of some agent with a completely different psychology from our own. (We surely can't even imagine what it's like to be Mary*.) Rather, our intuitions are shaped by our own psychological restrictions. And the physicalist who accepts my arguments can respect *these* intuitions.

While type-B physicalism fails, the result is a way of being a type-A physicalist that respects our pretheoretic intuitions. We do not need to, as Chalmers (2002) puts it, "deny the manifest", asserting that *if we only knew more*, we could bridge the gap. Any creature psychologically like us will find an epistemic and explanatory gap,[27] which they cannot bridge, no matter how much physical information they might gain.[28] This means that captive Mary would be fooled if presented with a blue banana, that no amount of physical information would render zombies inconceivable to us, that no amount of physical information could close the explanatory gap. While physicalists must give up on an ideal gap between the phenomenal and the physical, they don't have to give up on an unbridgeable *human* gap. And, plausibly, this is what we were after all along.

---

[27] Again, it's highly plausible that we are not wired up like Mary*: (i) Mary* has a very particular sort of neural architecture – both in the formatting of their physical and phenomenal concepts and in the programming to match them up. There's no reason to assume that we'd be wired up in this way. (ii) Given the peculiarity and complexity of this cognitive architecture, together the lack of any apparent evolutionary benefit to being designed like this, we have positive reason to suppose that we are not like Mary* in this respect. And (iii) the assumption that we are unlike Mary* in this both fits with our robust intuitions that we are unable to bridge the epistemic gap, and yields a elegant new version of type-A physicalism.

[28] One might wonder whether our ability to appreciate the existence of creatures like Mary* shows that we can bridge the epistemic gap. Alas, it does not. Much as there could be creatures – perhaps many animals – whose brains do not equip them to come to know mathematical truths, so our brains (arguably) are just not wired up as would be required to come to know phenomenal-physical truths a priori. We have ~~the cognitive capacities to~~ appreciate that others might have different cognitive architectures that enable them to know things a priori that we cannot. But this appreciation does not bestow these capacities on us, any more than neurally enhancing a mouse with the power to appreciate other animals' mathematical capacities thereby gives it the capacity to do calculus.

Bibliography

Balog, Katalin "Acquaintance and the mind-body problem." *The Mental, the Physical.* Eds. Hill & Gozzano. (Cambridge UP 2012a)

——— "In Defense of the Phenomenal Concept Strategy." *Philosophy and Phenomenological Research* (2012b)

Bigelow, John and Pargetter, Robert. "Acquaintance with Qualia." *There's Something About Mary*. Eds. Ludlow, Nagasawa, Stoljar. (MIT UP 2004)

Chalmers, David and Jackson, Frank. "Conceptual Analysis and Reductive Explanation." *Philosophical Review* 110. (2001)

Chalmers, David. "Consciousness and Its Place in Nature." *Philosophy of Mind: Classical and Contemporary Readings.* Ed. Chalmers (Oxford, 2002)

——— "The Content and Epistemology of Phenomenal Belief." *Consciousness: New Philosophical Perspectives*. Eds. Smith & Jokic. (Oxford UP 2003)

——— "Does Conceivability Entail Possibility?" *Conceivability and Possibility.* Eds. Gendler & Hawthorne. (Oxford UP 2002)

——— "Materialism and the Metaphysics of Modality." *Philosophy and Phenomenological Research* 59. (1999)

——— "Phenomenal Concepts and the Explanatory Gap." *Phenomenal Knowledge and Phenomenal Concepts.* Eds. Alter and Walter (Oxford UP 2007)

——— "Two-Dimensional Argument Against Materialism." *The Character of Consciousness.* (Oxford UP 2010)

Groh, Jennifer. *Making Space: How the Brain Knows Where Things Are.* (Harvard UP: 2014)

Jackson, Frank. *From Metaphysics to Ethics.* (Oxford UP: 2000)

Kripke, Saul. *Naming and Necessity.* (Blackwell 1972)

Levin, Janet. "What is a Phenomenal Concept?" *Phenomenal Knowledge and Phenomenal Concepts.* Eds. Alter & Walter (Oxford UP 2007)

Levine, Joseph. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical*

Quarterly 64. (1983)

——— "On Leaving Out What It Is Like." *Consciousness: Psychological and Philosophical Essays.* Eds. Davies & Humphreys (Blackwell 1993)

Loar, Brian. "Phenomenal States (Revised Version)" *The Nature of Consciousness*. Eds. Block, Flanagan, Güzeldere (MIT UP 1997)

Perry, John. "Frege on Demonstratives." *Philosophical Review* 86. (1977)

——— *Knowledge, Possibility, and Consciousness.* (MIT UP 2001)

Papineau, David. *Thinking about Consciousness*. (Oxford UP 2002)

——— "Phenomenal and Perceptual Concepts." *Phenomenal Knowledge and Phenomenal Concepts.* Eds. Alter & Walter (Oxford UP 2007)

——— "What Exactly is the Explanatory Gap?" *Philosophia* 39. (2011)

Stoljar, Daniel. "Physicalism and Phenomenal Concepts." *Mind and Language* 20. (2005)

Strawson, Galen. "Realistic Monism: Why Physicalism Entails Panpsychism." *Journal of Consciousness Studies* 13. (2006)

Tye, Michael. "A Theory of Phenomenal Concepts." *Minds and Persons*. Ed. O'Hear. (Cambridge UP 2003)