Phenomenal Knowledge *Why*: The Explanatory Knowledge Argument against Physicalism

Hedda Hassel Mørch
The Knowledge Argument: Then and Now (ed. Sam Coleman)

Penultimate draft—please cite published version.

1 Introduction

Phenomenal knowledge is knowledge of *what it is like* to be in conscious states, such as seeing red or being in pain. According to the knowledge argument (Jackson 1982, 1986), phenomenal knowledge is knowledge *that*, i.e., knowledge of phenomenal facts. According to the ability hypothesis (Nemirow 1979; Lewis 1983), phenomenal knowledge is mere practical knowledge *how*, i.e., the mere possession of abilities. However, some phenomenal knowledge also seems to be knowledge *why*, i.e., knowledge of explanatory facts. For example, someone who has just experienced pain for the first time learns not only *that* this is what pain is like, but also *why* people tend to avoid it.

Some philosophers have claimed that experiencing pain gives knowledge why in a normative sense: it tells us why pain is bad and why inflicting it is wrong (Kahane 2010). But phenomenal knowledge seems to explain not (only) why people *should* avoid pain, but why they *in fact* tend to do so. In this paper, I will explicate and defend a precise version of this claim and use it as a basis for a new version of the knowledge argument, which I call *the explanatory knowledge argument.* According to the argument, some phenomenal knowledge (1) explains regularities in a distinctive, ultimate or regress-ending way, and (2) predict them without induction. No physical knowledge explains and predicts regularities in the same way. This implies the existence of distinctive, phenomenal explanatory facts, which cannot be identified with physical facts.

I will show that this argument can be defended against the main objections to the original knowledge argument, the ability hypothesis and the phenomenal concept strategy, even if it turns out that the original cannot. In this way, the explanatory knowledge argument further strengthens the case against physicalism.

2 Background and Overview

The knowledge argument (Jackson 1982, 1986) is based on the thought experiment of Mary the color scientist. Mary is a gifted scientist who grows up in a room where everything is black and white. Here she has obtained complete physical knowledge about human color vision from black and white television. She is then released from the room into the world, and for the first time she sees a colored object, a ripe tomato. It seems then she will learn something new: she will obtain phenomenal knowledge about what it is like to see red.

The thought experiment could also be put in terms of pain. Suppose Mary has complete physical knowledge about the physiology of pain, but has never actually experienced it. Perhaps the black and white room is very safe and Mary has perfect health, so she has never had an accident or suffered from any kind of illness. Or, we might suppose Mary suffers from congenital insensitivity to pain, a medical condition which renders sufferers incapable of experiencing bodily pain of any kind. As the brilliant scientist she is, she figures out a cure for this condition which she applies to herself. Then she has an accident, she burns her hand, and experiences pain for the first time. It seems she will then learn something new: she now has phenomenal knowledge of what it is like to be in pain.

According to the knowledge argument, phenomenal knowledge is knowledge *that* this is what seeing red or being in pain is like, i.e., knowledge of phenomenal *facts* (where facts are understood as ontological or non-propositional items). Furthermore, because it would be new to someone like Mary who already knows all the physical facts, phenomenal knowledge must be about non-physical facts. The existence of non-physical facts refutes physicalism, the view that the physical facts are all the facts. The argument can be summed up as follows:

1. All physical facts are knowable without experience.
2. Some phenomenal facts are not knowable without experience.
3. Therefore, some facts are non-physical.

By knowability without experience, I mean knowability within a black-and-white, pain-proof and otherwise experience-restricted room, or more generally, knowability without reliance on any *particular kind* of phenomenal experience.

In response to the knowledge argument, some physicalists have disputed that phenomenal knowledge is factual. According to Lewis' ability hypothesis (1983), phenomenal knowledge is mere practical knowledge, or knowledge *how*. When Mary sees a red object for the first time, she merely learns how to imagine, remember and recognize the physical state of having her retinas stimulated by a certain wavelength of light. Similarly, when she experiences pain for the first time, as in the alternative version of the scenario, she merely learns how to recognize tissue damage and other harmful bodily states, how to imagine and remember these states, and so on. Given the assumption that gaining new abilities does not require becoming aware of any new facts, the ability hypothesis would avert the threat phenomenal knowledge poses to physicalism.

Other physicalists, such as Loar (1997) and Papineau (2002), grant that phenomenal knowledge is factual, but dispute that it is about any *new* facts. Rather, they claim, phenomenal knowledge is about the same old physical facts that someone like Mary would already know. When Mary experiences color or pain for the first time, she would learns to represent or conceive of known physical facts in a new and different way. This response is known as the phenomenal concept strategy. If phenomenal knowledge is about wholly physical facts, as per this response, it would also pose no threat to physicalism.

These are the main accounts of phenomenal knowledge, but there are also other variations. According to Kahane (2010), some phenomenal knowledge also constitutes a special kind of knowledge *why*, in the following sense: someone who experiences pain for the first time will learn not only *that* this is what pain is like, but also why pain is bad and why we should not inflict it on others. Phenomenal knowledge of pain thereby constitutes *normative* knowledge. Kahane also notes that no physical knowledge seems normative in the same way. He suggests that this could form the basis for a new argument against physicalism, a *normative knowledge argument*, although he does not go on to develop such an argument. [1]

In this paper, I will defend the claim that some phenomenal knowledge, of pain in particular, constitutes knowledge *why*, but in a factual rather than a (merely) normative sense: knowledge of what pain is like tells us not (only) why we *should* avoid pain, morally or rationally speaking, but (also) why people *in fact* tend to try to avoid it. That people generally try to avoid pain is an ordinary, empirical psychological regularity, not a normative claim (though it is of course compatible with the normative claim). I will argue that phenomenal knowledge of pain (1) *explains* this regularity in a distinctive, ultimate or regress-ending way, and (2) *predicts* it without induction, but no physical knowledge explains and predicts this, or any other, regularities in the same way. Furthermore, these distinctive explanatory features of phenomenal knowledge reflect distinctive explanatory *facts.* This gives the basis for what I will call *the explanatory knowledge argument*:

1. All physical facts are knowable without experience
2. Some explanatory facts are not knowable without experience.
3. Therefore, some facts are non-physical.

What kinds of facts are explanatory facts? I will claim that, in this case, they are facts about *causal powers.* That is, phenomenal knowledge of pain is distinctively explanatory and predictive because pain itself seems to have the power to make subjects who experience it try to avoid it, and it appears to have this power *in virtue of how it feels*, or its phenomenal character. And given that no physical knowledge enables the explanation and prediction of any regularities, there do not seem to be any physical causal powers of the same sort.

In what follows, I will articulate and defend these claims in more detail (section 3). I will then consider a number of objections, including the objection that physical knowledge can be equally explanatory as phenomenal knowledge given dispositional essentialism (the view that physical properties are essentially dispositional or powerful) (section 4), and objections based on apparent exceptions to the regularity between pain and avoidance attempts (section 5)—such as the medical condition *pain asymbolia*, where patients report feeling pain that they have no inclination to avoid (Grahek 2007).

[1] Instead, he develops a normative knowledge argument against externalism in metaethics.

I will then argue that the explanatory knowledge argument is resistant to both the ability hypothesis and the phenomenal concept strategy, in ways the original is not (section 6), and thereby strengthens the case against physicalism relative to the original knowledge argument. The reason for this is roughly as follows. The original knowledge argument claims that phenomenal knowledge would be simply *new* to someone like Mary. In response, the ability hypothesis and phenomenal concept strategy claim that this knowledge is not about any new *facts*, but can rather be explained away in terms of new *abilities* or *concepts*. The explanatory knowledge argument, in contrast, claims that some phenomenal knowledge would be new to Mary *in virtue of* being distinctively explanatory. To explain how phenomenal knowledge could be distinctively explanatory, the ability hypothesis and the phenomenal strategy would have to posit not only new, but also distinctively explanatory abilities or concepts. Butas I will argue, it is not clear how abilities or concepts could be explanatory if there are no corresponding explanatory facts (such as causal powers).

After a preliminary summary (section 7), I will then consider how the explanatory knowledge argument relates to a potential normative knowledge argument, as suggested by Kahane (section 8). Finally (section 9), I will consider some further implications of the explanatory knowledge argument for mental causation and the principle of physical causal-explanatory closure.

3 The Explanatory Knowledge Argument

To repeat, the explanatory knowledge argument goes as follows:

1. All physical facts are knowable without experience (i.e., within a black-and-white, pain-proof and otherwise experience-restricted room).
2. Some explanatory facts are not knowable without experience.
3. Therefore, some facts are non-physical.

Premise 1 of this argument overlaps with premise 1 of the original knowledge argument. It is rarely disputed (some even take it as true by definition), and I will therefore take it for granted. Premise 2 will be defended by appeal to the following sub-argument:

1. No knowledge available without experience (i.e. no physical knowledge) (1) ultimately explains regularities and (2) predicts them without induction.
2. Some knowledge available from experience (i.e., phenomenal knowledge) (1) ultimately explains regularities and (2) predicts them without induction.
3. Knowledge that (1) ultimately explains regularities and (2) predicts them without induction is about explanatory facts.
4. Therefore, some explanatory facts are not knowable without experience.

I will assume that knowledge available without experience is in principle exhausted by physical knowledge, by which I mean knowledge of (ideal/completed) physics (or the non-mental empirical sciences) and knowledge that can in principle be deduced from it. I will also take physical *facts* to

be exhausted by the kinds of facts that can (in principle) be completely described by physical knowledge. [2]

Knowledge available from experience includes phenomenal knowledge. My argument will presuppose the existence of phenomenal knowledge,[3] but to be clear, it will not presuppose that phenomenal knowledge is either factual or about any non-physical facts (rather, this is what the argument aims to establish, and it will also be defended against the ability hypothesis and the phenomenal concept strategy).

I will now defend each premise of the supporting argument in turn.

3.1 Premise 1: No Physical Explanatory Knowledge

The first premise of the supporting argument first claims that no knowledge available without experience, which (as noted) I take to be equal to physical knowledge, *ultimately explains* regularities. By an ultimate explanation, I mean an explanation that does not give rise to further why-questions, because it does not appeal to anything contingent or inexplicable, but rather to something that is itself necessary, self-evident or self-explanatory. Some putative examples of ultimate explanations, outside the realm of the physical and phenomenal, include mathematical explanations that appeal to self-evident axioms and theological explanations that appeal to God understood as a necessary being.

By *regularities,* I mean lawlike generalizations, including physical laws, laws of special sciences, and behavioral or cognitive regularities that could be considered laws of psychology. To count as a regularity, a generalization must hold invariably true in the absence of interference, or conflict with other regularities, i.e., *ceteris absentibus*.

It is fairly clear that no physical knowledge ultimately explains any regularities. Some regularities can be physically explained in a non-ultimate way, e.g., laws or regularities of physiology might be explained in terms of laws of chemistry, and laws of chemistry can be explained in term of the laws of physics. But the fundamental laws of physics cannot be explained—they are matters of brute, empirical fact. When asked why they hold, physicists would either say that we do not know, or that they just *do*—there is no explanation. Of course, it may turn out that the laws of current physics can be explained in terms of a more fundamental theory, such as string theory or multiverse

[2] Some might be skeptical to defining physical knowledge (and facts) in terms of physics, in view of, for example, Hempel's dilemma (according to which current physics is false but ideal physics is unknowable) or because one takes some special sciences to be autonomous but still physical. One might therefore rather define physical knowledge negatively in terms of what it is not. On the negative part of my definition, I follow Papineau (2001) and Wilson (2006). Note that some define physical knowledge more broadly as knowledge of the nature of the kinds of *objects* described by physics, and leave it open whether physics (or the non-mental empirical sciences) can describe the full nature of these objects (see, e.g., Stoljar on "o-physicalism" (2001) and Chalmers on broad physicalism (2003)). Neither the original nor the explanatory knowledge argument are aimed to refute the kind of physicalism that takes all facts to be physical only in this broader sense. As will be discussed later, the explanatory knowledge argument also positively supports a view that may be classified as physicalism in this broader sense, namely Russellian monism.

[3] The existence of phenomenal knowledge is accepted by most physicalists, with the exception of extreme forms of eliminativism or illusionism. In this paper, I set these views aside.

theory. But these explanations would then depend on the laws of string theory or the laws of the multiverse, and there would be no explanation of why these laws hold.

To say that the fundamental laws of physics have no ultimate explanation is also to say that there is no physical knowledge in virtue of which they seem *necessary*. Any fundamental law of physics could conceivably be different, even given expert physical knowledge. Cosmologists, for example, often consider how the laws of physics could be different (e.g., more or less "fine-tuned" for life), as well as hypotheses according to which they actually are different (e.g., in different universes within a multiverse).

The first premise also claims that no physical knowledge *predicts* regularities *without induction*. Explanation and prediction are closely related by the fact that explanatory hypotheses usually predict the facts they explain. If regularities can be ultimately explained in terms of something else than other regularities, one would expect them to be predictable based on this explanation alone, as opposed to on the basis of induction from multiple observations, which is our usual tool for discovering regularities.

It is widely agreed that no physical regularities can be predicted without induction. Sometimes, they can be predicted without induction being *directly* involved, as when regularities of higher level sciences are deduced from the underlying laws of physics. But the laws of physics must then already have been confirmed inductively.

3.2 Premise 2: Phenomenal Explanatory Knowledge

The second premise of the supporting argument claims that some phenomenal knowledge ultimately explains regularities and predicts them without induction. How could this be? Clearly, no phenomenal knowledge can ultimately explain or non-inductively predict the laws of physics.[4] But consider the psychological regularity "pain makes all subjects who experience it try to avoid it". This regularity seems to hold true in the absence of interference from other motives or reasons, i.e., *ceteris absentibus*. It is of course true people often endure or pursue pain for various kinds of interfering motives: some endure pain because they believe it will lead to less pain in the future (as when cleaning a wound), some pursue pain because the pain is accompanied by pleasure (as in masochism), some endure pain because they believe it is morally appropriate (as when accepting punishment). But in the absence of any further motives, it seems people (and other animals, as far as we can tell) always try to avoid it, i.e., we never endure or pursue pain for absolutely no reason.

The regularity also does not seem to positively depend on further beliefs about pain, such as that pain is dangerous—otherwise we would not take painkillers for knowingly harmless headaches.[5] Nor does it seem to depend on contingent attitudes such as fear of the pain (although fear could

[4] Unless panpsychism is true. As will be discussed below, panpsychism (of the Russellian monist kind) is one of the non-physicalist views compatible with the explanatory knowledge argument.

[5] Some (e.g., Cutter and Tye 2014) try to explain why it is rational to take painkillers in other ways, but intuitively, we do it in order to avoid the phenomenal experience of the pain itself. As will be discussed below, the motivational power of pain might depend on beliefs agents have about themselves, but it does not seem to depend on beliefs about the pain.

cause the pain to get worse, or constitute an additional motive that makes us even more inclined to try to avoid it).

It should also be noted that the regularity I will consider only holds between pain and *tryings*, i.e., efforts or attempts, to avoid it, where these tryings should be understood as purely mental events.[6] A further regularity seems to hold (again, in the absence of interference) between efforts to avoid pain and actual, successful avoidance (or between tryings and successful actions in general), but this regularity is distinct from the regularity between pain and mere tryings to avoid it, so to explain one is not necessarily to explain the other.

Why does the regularity between pain and avoidance attempts hold? Consider the overprotected or congenitally insensitive Mary, who has complete physical knowledge about pain, but has never experienced it. She leaves the room, cured of any insensitivity, and has her first accident—she badly burns her hand. Upon having this experience, it seems she would not only think: "Aha, so this is what it is like be in pain!", but also: "I now understand *why* people try to avoid it."

Knowing how pain feels, it seems self-explanatory why people try to avoid it. People avoid pain *because it feels like this*. This explanation invokes no further regularities, only the intrinsic character of pain. Furthermore, it gives rise to no regress of further why-questions. When we explain a law of chemistry in terms of a law of physics, we can ask: "but why do the laws of physics hold?"—and get no answer. But when Mary understands that people try to avoid pain because it feels like *this* (when pointing to her own experience), she would not ask the further question: "but why does something that feels like *this* make people avoid it?" This can be answered simply by attending to the phenomenal character of pain again.

Knowing how pain feels, it is also hard to conceive of this regularity being otherwise, especially the scenario of strong pain and pleasure being inverted in our motivational structure. Could intense, terrible pain, in and of itself, in the absence of any interfering motives, make us try to have more of it? Could intense, blissful pleasure, in and of itself, make us try to avoid it?[7] For someone who has never experienced either pain or pleasure, this would be just as conceivable as different laws of physics. But once we think of pain and pleasure in terms of how they feel, i.e., take their phenomenal character into consideration, it is very hard to imagine.

Knowledge of pain also seems to enable prediction of regularities without induction. This can be illustrated by another thought experiment. Imagine someone, call her Maya, who has also never experienced pain, and has not been as well educated as Mary: she does not know that pain makes subjects try to avoid it. Maya has some physical knowledge about pain physiology, such as that there is some bodily state that is correlated with something people call (phenomenal) pain. But she

[6] Note that presupposing the existence of mental tryings does not beg the question against physicalism. Physicalists (except eliminativists) generally accept the existence of mental events such as tryings, they just regard them as identical with or constituted by physical events.

[7] One might think very intense or prolonged pleasure can get uncomfortable or boring and therefore eventually make us try to avoid it. But if so, it would seem that either the phenomenology will have changed into something that no longer feels like pleasure, or the discomfort or boredom would constitute a distinct, interfering motive.

does not know that this bodily state causes avoidance attempts, nor does she know any lower level physical regularities from which this could be deduced. Then she experiences pain for the first time and learns what it is like, say by stepping on a sharp nail (barefoot, in a way that feels absolutely horrible). It seems she would then be in a position to instantly predict that this is a feeling she, and everyone else who experiences it, will try to avoid in the future, unless they have a further reason not to. She would not need to observe her own reaction to pain multiple times, and observe the same reaction in others, and then apply inductive reasoning. Rather, she could predict it from a single experience of pain alone.

At this point, one might object that this apparent explanatory and predictive knowledge may be illusory. In particular, one might object that it is not truly *inconceivable* that pain does not make a subject try to avoid it (*ceteris absentibus*). If this is not truly inconceivable, neither would it truly seem necessary and self-explanatory, and thus ultimately explained. It would also undermine the claim that phenomenal knowledge (alone) enables non-inductive prediction, because if prediction from phenomenal knowledge is not based on inconceivability, it seems it would rather have to be based on some additional, implicit associations or assumptions.

In response, I will now attempt to demonstrate that it is truly inconceivable that pain and avoidance attempts come apart in view of phenomenal knowledge, *given certain qualifications*. I will then argue that this gives reason to suppose that phenomenal knowledge reflects explanatory facts, in the form of phenomenal causal powers—as per premise 3, the final premise of the supporting argument. If phenomenal explanatory knowledge is about such explanatory facts, it would also be veridical and non-illusory. My defense of premise 3 will thereby also serves the purpose of answering the objection.

3.3 Premise 3: Phenomenal Explanatory Facts

Is it truly inconceivable that pain does not make a subject who experiences it try to avoid it (*ceteris absentibus*)? It might be conceivable that pain is not regularly *followed* by avoidance attempts, as per the regularity theory of causation (Hume 1739; Lewis 1973). It might also be conceivable that pain is necessarily connected to something else than avoidance attempts in virtue of external governing laws or relations (Dretske 1977; Tooley 1977; Armstrong 1978); or that pain has no effects at all, as per epiphenomenalism (Jackson 1982).

What does not seem conceivable, however, is that pain *makes* us try to pursue it, remain indifferent to it, or otherwise do anything else than avoid it, in virtue of its intrinsic, phenomenal character alone. Or conversely, that pleasure *makes* us try to do anything else than pursue it, in virtue of its respective intrinsic, phenomenal character alone. That is to say, assuming causation is a matter of non-Humean *production*, and that pain and pleasure produce their effects *in virtue of how they feel*, it seems inconceivable that they produce different effects than their actual ones. After all, is not the phenomenal character of pain just intrinsically *disagreeable* and *repulsive*, and the phenomenal character of pleasure not just intrinsically *agreeable* and *attractive*? So, if pain and pleasure produce their effects in virtue of these respective qualities, how could they possibly make subjects respond otherwise?

I will now consider this conditional inconceivability claim in more detail. Does it really hold in anything but a trivial sense? And even if it does, would not appealing to it in defense of premises 2 and 3 beg the question by presupposing the existence of productive causal powers, i.e., explanatory facts in virtue of which phenomenal explanatory knowledge is factual and thus veridical (as per premises 3 and 2 respectively)?

To repeat, the conditional inconceivability claim is claim that it is inconceivable that pain and avoidance attempts come apart assuming that pain has causal powers in virtue of how it feels (as opposed to being causally relevant in virtue of external regularities or governing laws, or having no causal relevance at all). Causal powers can be defined, more precisely, as properties in virtue of which causes metaphysically necessitate their effects (in the absence of interference from other powers, i.e., *ceteris absentibus*) by *producing* them, or *making* them happen. To say that pain has causal powers in virtue of how it feels is therefore to say that the phenomenal properties of pain metaphysically necessitate their effects in this way.

One might think that it is inconceivable that pain has different powers or necessitate different effects in virtue how of it feels only because it is an analytic truth that *if* pain has causal powers, then it necessitates its actual effects. But the inconceivability only depends on accepting that pain has causal power in a general sense. Assuming pain has *some* causal power in virtue of how it feels, it is inconceivable (considering how it feels) that it should have anything other than the *particular* power to make subject try to avoid it. It is does not follow trivially from "pain has *some* power in virtue of how it feels" that it has any particular power, or that it could not have *different* powers or effects. This should be clear from the fact that the same result does not follow from assuming that physical objects have causal powers. Assuming physical objects have *some* causal powers, it is still conceivable that they have different particular powers or effects than those they actually have. For example, assuming billiard balls have some causal powers, it is still conceivable that they have the power to pass through other objects on impact, or jump over them, and so on.

One might object that the inconceivability must nevertheless be based on implicitly assuming that pain necessitates not just *some* effect or other, but its *particular*, actual effects, as per some form of analytic functionalism. According to analytic functionalism, the concept of pain just is the concept of having some particular functional or dispositional role, such as making subjects try to avoid it. If pain is conceived of in this way, it would be an implicit logical contradiction to say that the regularity between pain and avoidance attempts does not hold. But in that case, phenomenal knowledge will not have succeeded in ultimately explaining any *causal* regularities. Rather, it would at best have succeeded in explaining what may be construed either a mere *analytic*, non-empirical truth or as a mere *constitutive* relation between a functional or dispositional property (i.e., pain functionally understood) and its constitutive input and output (i.e., the output of avoidance attempts given the input of being experienced by a subject)—both of which are things that physical knowledge would also clearly be capable of explaining in the same way.

But the inconceivability does not depend on conceiving of pain in functional or dispositional terms. In order to render it inconceivable that pain produces different effects in virtue of how it feels, it is

sufficient to conceive of pain in terms of a demonstrative, phenomenal concept ("feeling like *this*"). There is no logical contradiction implicit in the claim that something that feels like *this* (when pointing to an experience or vivid memory of pain) fails to make (in the non-Humean sense) a subject try to avoid it (*ceteris absentibus*).[8] But it still seems inconceivable.

If the conditional inconceivability claim could thus be shown to non-trivially hold, it might nevertheless seem question-begging to assume the correctness of antecedent, i.e., that pain has causal power in virtue of how it feels, or that phenomenal properties *produce* their effects. To avoid this charge, the antecedent assumption could be supported by arguments for realism about causal powers that are independent of the explanatory knowledge argument.[9] But in that case, the explanatory knowledge argument would not really constitute an argument against physicalism, but rather an argument that realism about causal powers is incompatible with physicalism. This conclusion would still be highly significant, given that realism about causal powers is a widely held view and generally regarded as perfectly compatible with physicalism. But the conditional inconceivability claim might also be able to support a stronger argument against physicalism as such, without either begging the question or invoking any separate arguments—because there is a sense in which the conditional inconceivability claim constitutes evidence for its own antecedent. How could this work?

First of all, if it is inconceivable that pain has different effects assuming that it has *some* causal power, it means we at least understand how pain *could* productively necessitate its actual effects. In other words, the conditional inconceivability claim shows that it is *positively conceivable* (i.e., imaginable in qualitative detail) how pain could have causal powers in virtue of how it feels, and positive conceivability is strong evidence of possibility.

For physical properties, in contrast, it may be *negatively conceivable* (i.e., not involve explicit or implicit logical contradiction) that they necessitate their actual effects. But it is not positively conceivable, as it is for pain, given that we can equally well conceive of physical properties or

[8] The phenomenal concept of pain might still be *conceptually* connected to avoidance attempts in a broader, non-logical, sense. Chalmers proposes that pure phenomenal concepts are constituted by the phenomenal properties they refer to (or "faint Humean copies" thereof) (Chalmers 2010b: 265-266, 272). If there is a necessary connection between phenomenal pain and avoidance attempts, and the concept of pain is constituted by (a Humean copy of) pain, then there will also be a necessary connection between the concept of pain and avoidance attempts. But this connection would not obtain in virtue of a constitutive relation between the concept of pain and the concept of avoidance attempts, but rather in virtue of a causal connection between the non-conceptual constituents/referents of these concepts.

Also note that Chalmers distinguishes phenomenal concepts from ordinary demonstrative concepts, because demonstrative concepts standardly leave the nature of their referent open (their meaning, in the phenomenal case, could be glossed as "this quality, whatever it happens to be") (Chalmers 2010a: 258). In Chalmers' terms, therefore, the causal power of pain is only explicable in terms of a concept that is both demonstrative (in a broad sense) and qualitative, i.e., which does not leave the nature of its referent entirely open.

[9] Such as the argument that without causal powers, the regularities of the world would constitute an enormous "cosmic coincidence" (Strawson 1987), or that realism about causal powers is necessary to justify induction (Ellis 2010). One might think one would also have to appeal to arguments against epiphenomenalism, but this will be redundant as part of an argument against physicalism, because physicalism takes phenomenal properties to be physical and no physical properties are epiphenomenal.

objects necessitating any other effects, and positive conceivability is much stronger evidence for possibility that negative conceivability.

Strictly speaking, if realism about causal powers is merely possibly true for pain, this would be sufficient to refute physicalism. If phenomenal properties necessitate their effects in some possible worlds where realism about causal powers is true, but physical properties do not necessitate them in any worlds, this suffices to show that the two sorts of facts are not identical—since identities are necessarily true if true at all, and identical properties do not differ in any possible world.

However, it might be objected that a metaphysical view such as realism about causal powers must be necessarily true if true at all. It follows that if the view is not actually true, it cannot be possibly true either. Therefore, it is worth noting that the conditional inconceivability claim may also support a further, direct argument that realism about causal powers is *actually* true for pain.

As already argued, the conditional inconceivability of pain and avoidance attempts coming apart shows that we have a positive conception of how realism about causal powers could be true for pain. But this positive conception does not seem like a conception of a mere possibility conjured up by the imagination. Rather, it is a view that we naturally and intuitively adopt, in most cases implicitly, on the basis of experience. For example, going back to the thought experiment of Maya, when Maya experiences pain for the first time, it seems plausible that she would naturally and implicitly accept that the phenomenal character of pain determines its causal powers, because her experience would seem to present it as such (that is, her experience would not present it as being completely up in the air whether its causal relevance is rather determined by external regularities or laws, or whether it might have no causal relevance at all; rather, her experience would seem to positively suggest that pain has causal powers determined by its phenomenal character). Positive conceptions derived from experience, rather than the imagination, are generally regarded as *appearances*. If pain thereby appears powerful, this constitutes evidence that it actually is powerful.

To recap the argument so far: It seems inconceivable that pain *makes* subjects who experience it do anything else than avoid it, in virtue of how it feels (as opposed to in virtue of a governing law, or in virtue of contingent regularities). It follows that *if* pain has causal powers in virtue of how it feels, it necessitates avoidance attempts. This conditional necessity shows that we can positively conceive of how pain *could* necessitate avoidance attempts. Furthermore, because this positive conception is derived from experience (as opposed to pure imagination), it constitutes an appearance that it actually does necessitate avoidance attempts. If the appearance is veridical, it establishes premise 3 (and at the same time refutes the above-mentioned objection to premise 2), because if phenomenal pain properties necessitate their effects, this would constitute an explanatory fact.

At this point, it might be objected that the appearance of pain having causal powers need not be veridical. But in general, appearances are taken to be veridical unless they conflict with other appearances or with important theoretical considerations. In this case, there are no obvious conflicts with other appearances. One potential source of conflict would be if the regularity theory, realism

about laws, or epiphenomenalism appeared to be true for pain in some other way. But these views seem more theoretically motivated than motivated by direct appearances, at least for phenomenal properties.

As for conflicting theoretical considerations, one might argue that positing causal powers is unparsimonious. But in general, we do not consider appearances non-veridical simply because that would be more parsimonious—if we were to maximize parsimony in this way, we should consider every appearance non-veridical and embrace solipsistic external world skepticism. Also, given that accepting the appearance as veridical has explanatory value with respect to phenomenal regularities, theoretical considerations also speak in favor of it.

Therefore, even though the appearance of pain having causal power could coherently be dismissed as false, there is no obvious reason to dismiss it as false (except the question-begging reason that it would lead to a problem for physicalism). It is implausible to dismiss appearances as false without any (non-question-begging) reason. The burden of proof is therefore on physicalists to point out some further, less obvious reason to dismiss the appearance.

This concludes my main case for the supporting argument for the claim that some explanatory facts are not available without experience, i.e., without phenomenal knowledge. This claim constitutes the only controversial premise of the explanatory knowledge argument. I will now consider further objections to each premise of this supporting argument, before summarizing the entire defense.

4 Objections to Premise 1

4.1 Physical Causal Powers

In view of my defense of premises 2 and 3 of the supporting argument, according to which phenomenal knowledge is ultimately explanatory and non-inductively predictive *assuming* phenomenal properties involve causal powers (as they also appear to), one might have the following objection to premise 1: could not physical knowledge be explanatory and predictive in the same way assuming physical properties involve causal powers?

In particular, it might seem physical knowledge would be capable of this assuming dispositional essentialism (Shoemaker 1980; Mumford 2004; Bird 2007). Dispositional essentialism is the view that all properties are essentially dispositional or powerful (I will use these terms interchangeably). For example, the physical property of solidity would essentially consist in (roughly) the power to avoid spatial overlap with other solid objects. This may seem to ultimately explain the regularity "solid objects do not pass through each other", and render it inconceivable that it does not hold, because to say that solid objects pass through each other would be to say that they are not solid after all, i.e., it would be a contradiction in terms. In the same way, the property of having negative charge could be regarded as essentially consisting in (roughly) the power to repel other entities with negative charge and attract entities with positive charge. This explains the regularity "electrons repel other electrons" insofar as electrons are essentially negatively charged.

However, physical knowledge would not seem have the same kind of explanatory and predictive powers as phenomenal knowledge even given dispositional essentialism. This can be seen by the fact that explanations of regularities in terms of essential dispositions always seem to involve *analyticity*. Take solidity. We can think about this solidity either in terms of a dispositional concept (analyzable roughly as "the property of being disposed to not pass through other solid objects"), a categorical concept (such as the property of having certain qualitative look or feel) or a demonstrative concept ("*that* property"). If we conceive of solidity in dispositional terms, as "the property of not passing through other solid objects", it will be contradictory and inconceivable that solid objects pass through other solid objects, and thereby explicable and necessary that the regularity that they do not pass through each other holds. But if we conceive of solidity in terms of a non-dispositional, either categorical or demonstrative concept, it would no longer be inconceivable that solid objects pass through each other, and the regularity would not seem necessary or explicable. For this reason, physical knowledge of solidity only seems capable of explaining what may be regarded either as a mere analytic truth or as a mere constitutive relation between the disposition of solidity and its constitutive input and output (i.e., the output of not passing through other objects given the input of intersecting paths of motion).

In contrast, I have argued that phenomenal knowledge of pain can explain regularities even when pain is conceived of under phenomenal concepts which are neither functional nor dispositional.[10] Therefore, phenomenal knowledge seems capable of explaining what seems like a properly causal regularity.

It might seem that dispositional essentialism nevertheless enables prediction without induction of properly causal regularities from physical knowledge. Nancy Cartwright has argued that, in practice, scientists often generalize from single observations (Cartwright 1999: 85). She argues that this practice, which clearly seems legitimate, can only be justified on the assumption that powers (or *capacities*, in her terms) belong to the essential natures of things. Roughly, her claim is that if the behavior of things is assumed to derive from their intrinsic powerful natures, then it is possible for a single instance of behavior to serve as a reliable indicator of this nature.

But as Cartwright explicitly notes, the view that powers belong to the natures of things is only a necessary condition to warrant generalization from single instances, not a sufficient condition. In order to make sure that a given experimental observation actually reveals the true nature of, e.g., electrons, scientists need to make sure that there is no interference which stops this nature from manifesting. And to rule out interference, they need to rely on inductively confirmed background assumptions about everything from the behavior of the experimental equipment to the workings of gravity, background radiation and so on. Dispositional essentialism (or the related capacities view) thereby fails to fully warrant prediction without induction, because induction must be involved indirectly in supporting necessary background assumptions.

[10] As discussed above, phenomenal concepts have a demonstrative element, but they may also have a categorical, qualitative element (see footnote 8 above).

In contrast, in the thought experiment of Maya—who has never experienced pain and does not already know that it makes all subjects try to avoid it—her prediction does not seem to rely on any inductively justified background assumptions about the absence of interference. The only thing that can prevent pain from making us try to avoid it are our own interfering motives, and these can be directly detected by us without induction because they would be constituted by our own occurrent mental states.[11]

Furthermore, interference might not even matter in the phenomenal case. Consider a scenario where Maya experiences pain for the first time in the following way. Maya knows she is going to learn what pain is like by touching an electrocuting wire. She is determined to study the nature of pain because she has a passionate interest in phenomenology, and she has trained herself to have complete control over her reflexes so that she will not avoid it involuntarily. When she touches the wire, she endures the pain (until the current is turned off and the pain ends by itself) because her interest in studying pain constitutes an interfering reason. In this way, she will experience pain, but will not witness it actually making her try to avoid it. It seems she would nevertheless infer that pain *would* make her try to avoid it were it not for her determination to endure it. Phenomenal knowledge thereby also seems to predict regularities without induction *despite* interference.

5 Objections to Premise 2

I will now consider objections to my defense of premise 2, according to which phenomenal knowledge of pain ultimately explains and non-inductively predicts the regularity "pain makes subjects try to avoid it *ceteris absentibus*". This defense might face objections according to which this regularity does not really hold. As discussed above, the most obvious apparent exceptions to the regularity can be classified as instances of interference and thereby covered by the *ceteris absentibus* clause. But there are other apparent exceptions that may not be accounted for in this way.

5.1 Ability and Agency

First of all, it might seem that attempts to avoid pain can be prevented by physical inability, which cannot plausibly be regarded as interference. For example, a paralyzed person who can still feel pain would not be able to avoid pain. But they could still *try* to avoid pain, if they do not know that they are paralyzed. This trying would still be a real event, that would (given most forms of physicalism) physically correspond to the firing of some neurons in their brain. But what about a

[11] This is not to say that our own motives are always fully transparent to us. Often, we cannot accurately categorize our own motives based on non-inductive introspection alone. But it seems we are always in a position to detect the presence of *some* motivation or urge to avoid an action, even in cases when we do not know how to characterize this motivation more precisely (e.g., one might not know whether one is procrastinating out of laziness, anxiety, or something else, but one still knows that one somehow *feels like* not working).

Relatedly, one might wonder whether there could not also be interference from unconscious motives. If unconscious motives would only be detectible by induction, this would prevent prediction without induction and thereby undermine the argument. But the existence of unconscious motives could be accounted for as a matter of unconscious states indirectly affecting us by causing conscious but uncategorizable urges. If so, the absence of unconscious interference could be detectable without induction because the absence of the urges that signify them would be detectable without induction.

paralyzed person who knows they are paralyzed? Arguably, it is not possible to try to do something unless one believes one has at some minimal chance of succeeding. Someone who believes they have no physical capacity to avoid pain could still try to avoid it by some mental action, for example, by deliberately focusing on something else. But if this does not work, they might eventually come to believe they have no mental capacity to avoid pain either, and therefore stop trying. If this is right, one might have to say that phenomenal knowledge can explain and predict the more specific regularity that pain makes subjects try to avoid it *if* they believe that can avoid it. This is does not express a mere analytic truth or constitutive relation, and physical facts do not explain or predict this or any similarly qualified regularities, so it would still support the claim that phenomenal knowledge is distinctively explanatory.

Another potential exception to the regularity are subjects who have no power of agency at all. Strawson has argued that it is metaphysically possible for there to be conscious subjects who are not agents (1994: ch. 9). He defends the conceivability of sentient, intelligent creatures called the Weather Watchers, whose conscious life consists in observing and contemplating the weather, without ever trying to do anything about it.

There are a number of possible responses to this problem. One response is to claim that the Weather Watchers would still need the ability to act and try because observation still requires some form of active thinking, which by Strawson's own admission (2008: 231), seems to require some minimal form of "catalytic" agentive effort.

Another response would be to add another qualification. If non-agentive subjects are possible, then phenomenal knowledge would still explain and predict the regularity "pain makes all *agentive* subjects try to avoid it". Again, this does not express a mere analytic truth or constitutive relation—as (the concept of) trying to avoid pain is not constitutive of (the concept of) agency—and physical knowledge does not explain and predict this or any other similarly qualified regularities either.

A third response is to appeal to a deflationary notion of subjects, according to which subjects of experience are not independent substances, but rather "bundles" of phenomenal experiences standing in certain types of relations (these relations may be more substantive than those that are part of the very minimal Humean version, and maybe *sui generis*). If pain is causally powerful, and a given subject is a bundle that includes pain, it follows that this subject is also powerful and is thereby an agent. Given the deflationary view, then, the objection that there could be non-agentive subjects in pain is equivalent to the objection that the assumption that pain has causal powers may be false, and can be responded by appeal to the same arguments I have already offered in support of this assumption above (according to which, although this assumption may be coherently denied, it strongly appears to be true).

5.2 Pain Asymbolia

Another potential objection to the regularity between pain and avoidance is based on the phenomenon of pain asymbolia. Pain asymbolia is a medical disorder where patients report feeling pain that does not hurt, or that they have no inclination to avoid. According to the standard analysis

of this phenomenon, due to Nikola Grahek (2007), there is no good reason to doubt that this description of pain asymbolia is accurate, i.e., to think that asymbolics are wrong to categorize what they are experiencing as pain, or to think that they must have some hidden motive to resist their inclination to avoid it. If this is right, such cases would constitute a direct counterexample to the regularity I have argued phenomenal knowledge about pain explains.

But pain asymbolia is still compatible with a more specific phenomenal pain regularity. Even though asymbolics identify what they are feeling as pain, it nevertheless seems that asymbolic pain and normal pain *feel* different. They seem to be two different phenomenal experiences, which nevertheless have enough in common to both fall under the same general concept of pain.

This interpretation is supported by Grahek, who concludes that what pain asymbolia really shows is that normal pain, which appears as a simple and unified feeling, is really complex. Normal pain is a combination of two components: "On the one hand, there is pure pain sensation, and on the other hand, there is the pure feeling of unpleasantness, defying any further sensory specification" (Grahek 2007: 111). Furthermore, both components are phenomenal. Not only can the sensory component be experienced without unpleasantness, as in pain asymbolia. Unpleasantness is also a phenomenal quality (or a *feeling,* as described by Grahek above), and it is also possible to experience unpleasantness by itself. There are also reports of a condition opposite of pain asymbolia, where patients report having the experience of pure unpleasantness whose character they could not specify in any further detail (Grahek 2007: 108-111).

There is no evidence that people can experience *unpleasantness* phenomenology without trying to avoid it, so there still seems to be regularity between *phenomenologically normal*, i.e., non-asymbolic unpleasantness-including, pain and avoidance attempts.[12] Given that this kind of pain can also be picked out in phenomenal terms, i.e., by how it (or its unpleasant component) feels, and not merely in functional terms such as "any quality (or kind of pain) that makes subjects try to avoid it", this regularity does not reduce to a mere analytic truth or constitutive relation. The regularity seems ultimately explicable and non-inductively predictable based how phenomenologically normal pain feels in the same way I have argued regularities involving general (asymbolic or non-asymbolic) pain initially seems explicable and predictable based on how general pain feels. The case for the explanatory knowledge argument could therefore be reformulated with reference to phenomenologically normal (unpleasantness-involving) pain instead of general pain. Or more simply, I will (retroactively) stipulate that by pain I mean phenomenologically normal pain.

6 Objections to Premise 3

Finally, I will consider objections to my defense of premise 3, according to which phenomenal explanatory knowledge reflects explanatory facts in the form of causal powers. This defense may seem vulnerable to versions of the main objections to the original knowledge argument, the ability

[12] Does this imply that the sensory part of pain has no causal power? It could still have some other causal power than the power to motivate avoidance.

hypothesis and the phenomenal concepts strategy, according to which Mary's new phenomenal knowledge need not be accounted for in terms of any new phenomenal facts, but rather in terms of new abilities or new concepts for old, already known facts. Similarly, one might think Mary's (and Maya's) new *explanatory* knowledge (or the explanatory aspects of her new phenomenal knowledge) can be accounted for in terms of new abilities or concepts as opposed to explanatory facts.

These objections also highlight another potential problem for the explanatory knowledge argument, namely that it may seem dialectically redundant relative to original knowledge argument, because it might seem it could only be sound and defensible insofar as the original knowledge argument is also sound and defensible.

But as I will now argue, the explanatory knowledge argument is resistant to the ability hypothesis and the phenomenal concept strategy, even if it were to turn out that the original is not. This is roughly because even if new abilities or new concepts may explain why phenomenal knowledge seems to present *new* facts, they do not explain why it seems to present *explanatory* facts. For that one would have to posit explanatory abilities or explanatory concepts. But it is hard to see how abilities and concepts can be explanatory without also involving explanatory facts. In this way, the explanatory argument is more defensible than the original in at least some respects.

6.1 The Ability Hypothesis

According to the ability hypothesis (Lewis 1983; Nemirow 1979), Mary's new knowledge of red consist in abilities such as to recognize, remember and imagine physical facts. If Mary's (and Maya's) new knowledge of pain is explanatory and predictive, then the ability hypothesis could be expanded to say that some phenomenal knowledge also consists in abilities to explain and predict physical facts.

In fact, Lewis acknowledges that some phenomenal knowledge is predictive, and proposes to account for this precisely in terms of a predictive ability:

> … knowing what it's like is the possession of abilities: abilities to recognize, abilities to imagine, *abilities to predict one's behavior* by means of imaginative experiments. (Someone who knows what it's like to taste Vegemite can easily and reliably predict whether he would eat a second helping of Vegemite ice cream.) (Lewis 1983: 131, my emphasis)

One response to this predictive ability hypothesis would be that it is not plausible epistemologically speaking. Usually, facts can only be reliably predicted on the basis of other facts. How could facts about regularities involving pain (or Vegemite) be reliably predicted without any factual basis?

But this could perhaps be accounted for by an evolutionary hypothesis. For example, one might think it was fitness-enhancing for our ancestors to be able to anticipate the effects of pain prior to repeated experience, and those who happened to innately associate them would therefore be selected for. This could be supported by the fact that there is already evidence of other predictive abilities of this sort. For example, there is evidence that infants expect various principles of mechanical causation to hold, such as "no action at a distance" (Michotte 1963; Spelke et al. 1995),

which can be taken to suggest an innate association implanted by evolution. One might think a similar kind of bias is responsible for prediction of the pain regularity.

But even if this hypothesis could account for the predictive aspect of phenomenal knowledge, it could not account for the explanatory aspect. First of all, other psychological biases do not involve any sense of understanding or intelligibility, as in the pain case. For example, when we think about it, we do not discover any apparent reason why action at a distance would be impossible. Second, no known psychological biases render whatever we are biased against altogether inconceivable, as in the pain case. Action at a distance, for example, would be highly unexpected, intuitively strike us as implausible, and so on, but we can still conceive of it if we try. Therefore, to explain away the explanatory features of phenomenal knowledge, one would have to come up with an additional "explanatory ability" to go with the predictive ability, and it is hard to see what kind of ability this could be.[13]

6.2 The Phenomenal Concept Strategy

According to the phenomenal concept strategy, phenomenal knowledge is factual, but it is not about any new facts. When Mary learns what it is like to see red, she merely learns to conceive of the physical facts in a new way: she acquires a new phenomenal concept for a fact that she already knows via a physical concept or physical mode of presentation.

If Mary's (and Maya's) new knowledge of pain is explanatory and predictive, then the phenomenal concept strategy would have to be expanded to say that some phenomenal concepts are not only new and different from ordinary physical concepts, they are also distinctively explanatory, or capable of presenting the same old physical facts in a new, more explanatory way. I will consider the main versions of the strategy, which characterize phenomenal concepts in different ways, to see whether they could give rise to any kind of explanatoriness.

David Papineau (2002) argues that phenomenal concepts are *quotational*, which is to say that the concepts are constituted by instances or copies of the experiences they refer to. Using a phenomenal concept of pain therefore puts one in a distinct psychological state that activates a copy of pain itself, rather than just an abstract representation of it, and this will make it falsely appear as though there are phenomenal facts about pain that go beyond the physical facts (Papineau 2002: 170-171). In this way, quotational concepts would be new and different compared to non-quotational concepts, even if the facts they refer to are identical. But there is no clear sense in which quotational concepts would be more explanatory than non-quotational concepts, given that the facts they refer

[13] Physicalists could still coherently dismiss the apparent explanatory knowledge gained from experiencing pain as completely illusory—in the same way they could coherently dismiss apparent purely phenomenal knowledge (or the non-explanatory aspects of it) as completely illusory, as per eliminativism or illusionism. But as discussed above, it is implausible to dismiss appearances as illusory without offering some (non-question begging) reason. It is also implausible dismiss it on the basis of general skepticism about appearances in order to support that the appearance that pain involves causal powers is non-veridical, because this kind of skepticism would seem to overgeneralize to support solipsistic external world skepticism. In the same way, if physicalists invoke general skepticism about our feelings of intelligibility and understanding in order to undermine the explanatoriness of phenomenal knowledge, it could also risk undermining our claims to intelligibility and understanding in a wide range of other areas.

to are identical and thereby equally explanatory. That is, there is no reason to think activating a copy of a phenomenal property would be sufficient to make it appear explanatory, if the phenomenal property is in fact no more explanatory than a physical concept would reveal.

John Perry (2001) argues that phenomenal concepts are similar to *indexical* concepts. Indexical facts (e.g., "you are here") cannot be derived from non-indexical physical facts (e.g., a map) and thereby seem new to someone who knows all but only non-indexical facts, but indexical facts arguably do not pose any problem for physicalism. If phenomenal concepts were indexical, it might therefore explain how phenomenal facts also appear new. Could it also explain how phenomenal facts could appear explanatory?

Unlike quotational concepts, indexical concepts do have special explanatory properties. As Perry has also argued (1979), indexical concepts can be essential to explaining behavior. He illustrates this with the following scenario. Perry is following a trail of sugar around a supermarket in order to find and stop the shopper who is making a mess. He suddenly realizes that there is a torn sack of sugar in his own shopping cart, and thereby learns the indexical fact that *he* himself is making a mess. This realization explains why he stops looking for another shopper and starts rearranging the torn sack in his own cart. If he were to merely learn the non-indexical fact that John Perry is making a mess, this would not be sufficient to explain his behavior: he would also need to know that *he* (himself) is John Perry.

Phenomenal explanation of the pain regularity also involves the indexically individuated fact "pain feels like *this*". But the explanatory power of phenomenal facts goes beyond what can be accounted for by their indexical mode of presentation. One difference is that what explains Perry's behavior is his indexical *belief*, i.e., the fact that "Perry believes *he* (himself) is making a mess". The first-order indexical facts "*he* is making a mess" and "*he* is Perry" by themselves seem causally and explanatorily inert. In the pain case, in contrast, it is the first-order fact that "pain feels like *this*" that explains why subjects try to avoid it, not the subjects' indexical beliefs about the pain. If first-order indexically individuated facts are generally not distinctively explanatory, then one cannot see how phenomenal facts could derive their explanatory power from their indexicality alone.

Furthermore, Perry's indexical belief only explains his behavior given that he also has a *desire* to not make a mess. This gives rise to the further explanatory question "why do desires to X cause attempts at pursuing X", which does not seem to have any ultimate physical answer.[14] I have argued that the fact "pain feels like *this*" is sufficient[15] to ultimately explain why subjects try to avoid it

[14] It might have an ultimate analytic answer, if desires are individuated in terms of their psychological roles. It might also have a phenomenal answer in terms of how desires feel, as for pain, but this would further support the explanatory knowledge argument rather than physicalism.

[15] As discussed, phenomenal explanations might also require that subjects have the additional belief that they have some capacity to avoid the pain. But the same is true for explanations of behavior in terms of indexical beliefs, i.e., Perry must not only believe that he is making a mess, but also that he has a capacity to stop making a mess. Therefore, the precise difference would be that only phenomenal facts are sufficient to ultimately explain the behavior of subjects who believe they are capable of the behavior.

without any additional desire to not be in pain (or not have *this* kind of feeling).[16] This is another respect in which phenomenal facts are more explanatory than typical indexical facts.

Other versions hold that phenomenal concepts are *recognitional* (Loar 1997) or *perceptual* (Papineau 2006), but these versions of the strategy share the same problems. Recognitional concepts might be especially explanatory in virtue of involving indexicality, but as I have argued, indexical facts are not as explanatory as phenomenal facts. Perceptual concepts are, according to Papineau, like quotational concepts but without any demonstrative element (2006: 120). Like quotational concepts, then, they are new and different compared to non-quotational physical concepts, but not more explanatory. Or, to be clear, these concepts could very well be explanatory *because* they quote or reference distinctively explanatory phenomenal facts, but not in virtue of their quotational character alone.

Proponents of the phenomenal concept strategy might also consider the view that some physical facts have explanatory features that are only apparent when conceived under phenomenal concepts. The view that some physical facts are only explanatory when considered in terms of, if not phenomenal, then at least *intentional* (and hence mental), concepts is not unheard of. Davidson's anomalous monism (Davidson 1970/1980), an important form of non-reductive physicalism, puts this forth as a fundamental tenet. According to anomalous monism, some mental events, such as intentional actions, can only be explained by other mental events, such as beliefs and desires. Beliefs and desires cannot be type-identified with any physical events—there is no way of systematically deriving physical descriptions of events from their mental descriptions—but every mental event is token-identical with some physical event. However, when a mental event, such as a belief, is redescribed as a physical event (with which it is token-identical), it will no longer be explanatory of any mental events. A belief might explain an intentional action, but a brain state (or other physically described states or events) never will.

Could physicalists adopt the analogous view that some physical facts (or events) are only explanatory when considered under phenomenal concepts? It seems not. The reason anomalous monism still qualifies as a form of physicalism is that, according to the view, although physical events will not explain mental events, physical events *will* explain those physical events with which mental events can be token-identified. In other words, if a particular token mental explanandum is redescribed in physical terms, which is always possible in principle, then it will have some physical explanation. Thus, physical and mental events have the same explanatory properties with respect to the same events at the token-level, just not at the type-level.

According to the explanatory knowledge argument, no matter how you redescribe an explanandum such as "pain makes subjects that experience it try to avoid it", it will never have an ultimate physical explanation. As I have argued, no physical facts ultimately explain any regularities, and I

[16] One might object that subjects in pain necessarily desire not to be in pain. Maybe so, but if so it seems the desire would follow from and be explained in terms of how pain feels, so it would be compatible with pain providing an ultimate explanation.

take it that it is not possible to redescribe (in a way that renders it fit for explanation) a regularity as anything else than a regularity. Physical and phenomenal facts thus have different explanatory relations to the same facts also at the token-level. This cannot be regarded as compatible with physicalism.

7 Summary of the Argument

The explanatory knowledge argument claims that:

1. All physical facts are knowable without experience.
2. Some explanatory facts are not knowable without experience.
3. Therefore, some facts are non-physical.

The only highly controversial premise of this argument is premise 2, which I have defended by appeal to the following sub-argument:

1. No knowledge available without experience (i.e., physical knowledge) (1) ultimately explains regularities and (2) predicts them without induction.
2. Some knowledge available from experience (i.e., phenomenal knowledge) (1) ultimately explains regularities and (2) predicts them without induction.
3. Knowledge that (1) ultimately explains regularities and (2) predicts them without induction is about explanatory facts.
4. Therefore, some explanatory facts are not knowable without experience.

To support this sub-argument, I have argued that phenomenal knowledge about (phenomenologically normal) pain ultimately explains and non-inductively predicts regularities such as "pain makes subject try to avoid it *ceteris absentibus"* (or perhaps "pain makes *agentive* subjects, *who believe they have the capacity*, try to avoid it *ceteris absentibus"*). This is mainly supported by the thought experiments about Mary, who has complete physical knowledge but upon her first experience of pain gains new knowledge of why subjects try to avoid it, and Maya, who does not know that pain makes subjects try to avoid it but upon her first experience can immediately predict it.

I have also noted that these explanations and predictions may depend on the assumption that pain has causal power in virtue of its phenomenal character. I have argued that this conditional explanatoriness nevertheless shows that we can positively conceive of how its phenomenal properties *could* be explanatory in virtue of necessitating their particular effects. This positive conception seems derived from experience, not the imagination, and thereby constitutes an appearance that phenomenal properties actually necessitate their particular effects. Phenomenal properties thereby appear to involve causal powers, which would constitute explanatory facts.

One might claim that the appearance is illusory, but appearances should generally be accepted as veridical unless (1) they conflict with other appearances or important theoretical considerations, or

(2) there is a good explanation of how the appearance could arise despite not being veridical. But there are no obvious reasons of either kind to dismiss the appearance.

I have also argued that physical knowledge does not ultimately explain or non-inductively predict any properly causal regularities (as opposed to analytic truths or constitutive relations) even assuming that physical properties essentially involve causal powers, as per dispositional essentialism. Dispositional essentialism also does not enable non-inductive prediction of physical regularities, because even if scientists sometimes generalize from single experiments (which, according to Cartwright, would be legitimate assuming dispositional essentialism or the related capacities view), they always rely on inductively supported assumptions about the absence of interference. Such background assumptions are not necessary in the phenomenal case. The lack of physical explanatory and predictive knowledge implies the lack of physical explanatory facts, given that all physical facts can (in principle) be revealed by physical knowledge.

This concludes my case for the explanatory knowledge argument. I will now briefly discuss how, in view of this defense, it dialectically relates to the normative knowledge argument, and well as to further issues in philosophy of mind.

8 The Normative Vs. the Explanatory Knowledge Argument

As noted, Kahane suggests that one might construct a normative knowledge argument against physicalism, based on the claim that phenomenal knowledge uniquely explains why pain is bad. Such an argument would presumably look something like this:

1. All physical facts are knowable without experience.
2. Some normative facts are not knowable without experience.
3. Therefore, some facts are non-physical.

An apparent weakness of this argument is that physicalists could simply deny the existence of the normative facts in question, i.e., embrace moral or normative anti-realism. This problem is anticipated by Chalmers:

> Moral facts are not phenomena that force themselves on us. When it comes to the crunch, we can deny that moral facts exist at all... The same strategy cannot be taken for phenomenal properties, whose existence is forced upon us. (Chalmers 1996: 83-84)

Unlike the normative knowledge argument (henceforth NA), the explanatory knowledge argument (henceforth EA) does not presuppose moral realism (although it is compatible with it). This gives EA a dialectical advantage against physicalists who are prepared to reject moral realism (or at least realism about the moral disvalue of pain in particular).

One might object that EA still depends an analogous assumption of *explanatory* realism, in the form of realism about causal powers. I have shown how realism about causal powers can be defended on the basis of the apparent explanatoriness of phenomenal knowledge. Therefore, it cannot simply be rejected by physicalists without further argument. But perhaps moral realism can

be defended in analogous way. In response to Chalmers' objection on behalf of physicalism, Kahane briefly responds that:

> This is an odd remark. The badness of pain seems to force itself upon us just like phenomenal properties. Indeed it imposes itself on us through a phenomenal property! (Kahane 2010: 47, footnote 47)

This suggests that, in the same way realism about causal powers can be justified on the basis of how pain appears powerful, moral realism can be justified on the basis of how pain appears bad.

This would be good news for the case against physicalism, but one might think it would render EA dialectically superfluous relative to an equally strong NA. But this does not follow, because physicalists could still coherently (albeit implausibly, according to NA and EA) dismiss either of these appearances as non-veridical, and some physicalists might nevertheless find it harder to deny that pain is powerful than that pain is bad.

Furthermore, NA may be less resistant than EA to the ability hypothesis and the phenomenal concept strategy. Against NA, it could be suggested that normative knowledge of pain only consists in the possession of normative abilities or normative concepts for non-normative facts, as opposed to awareness of objective normative facts. There are some candidates for normative abilities or concepts, such as prescriptive or expressive abilities or concepts, that may seem potentially capable of explaining away the appearance of normative facts. If so, a normative ability hypothesis or normative concept strategy may be more plausible than an explanatory ability hypothesis or explanatory concept strategy.

9 Mental Causation and Physical Causal Closure

The explanatory knowledge argument is primarily an argument against physicalism, but it also has further implications for the metaphysics of phenomenal properties. First of all, the argument suggests that (some) phenomenal properties are not epiphenomenal, because they necessitate corresponding efforts. This would be a further difference between it and the original knowledge argument, which was first offered in defense of epiphenomenalism (Jackson 1982).

However, the argument does not strictly preclude epiphenomenalism, understood as the view that phenomenal properties have no *physical* effects, because epiphenomenalism leaves open the possibility that phenomenal properties have other non-physical effects, as long as these effects are also physically inert. I have argued that pain appears to necessitate *efforts* towards avoidance, where efforts are understood as purely mental events. As noted above, there might be a further regularity between these efforts and physical actions. But it is not as clearly inconceivable that efforts have different effects in virtue of their phenomenal (or otherwise intrinsic) character. Therefore, it seems coherent to hold that pain necessitates non-physical mental efforts, but that these efforts are in turn epiphenomenal with respect the physical world. On the other hand, it would also be compatible with the argument to posit a further, psychophysical regularity between mental efforts and physical actions in accordance with interactionism. But even though the argument is thereby be compatible with both epiphenomenalism and interactionism, it lends somewhat more to support interactionism because the kind of epiphenomenalism it implies (according to which not

all phenomenal properties are mentally inert even though they are all physically inert) seems even more inelegant than standard epiphenomenalism (according to which all phenomenal properties are both mentally and physically inert).

A third, and perhaps more natural, option, however, is to take phenomenal properties to be explanatorily related to the physical world in virtue of directly underlying *physical* regularities in accordance with what is known as Russellian monism. Russellian monism is the view that all physical properties are purely structural or relational, and that physical structure intrinsically realized by phenomenal or protophenomenal properties (i.e., properties that are neither physical nor phenomenal, but closely related to the phenomenal) (Alter and Nagasawa 2012; Chalmers 2013). This realization relation can be conceived of in different ways. One view is that physical structure is realized by non-powerful (proto)phenomenal properties related by contingent regularities or governing laws. But another possible view is that all physical structure is realized by (proto)phenomenal powers which in turn ground the regularities or laws. This possibility is further supported by the intelligible relation between phenomenal pain and pleasure and regularities emphasized by the explanatory knowledge argument. Russellian monism is thereby also compatible with, and to some extent supported by, the explanatory knowledge argument.

This compatibility is also relevant to another objection one might have against the argument, namely that it implies a vicious dilemma between epiphenomenalism and violation of the principle of physical explanatory closure, i.e., the principle that every physical event that has an explanation has a sufficient physical explanation. This principle is widely regarded as having strong empirical support (Papineau 2001). But the principle could be formulated in different ways. One version would be as follows: every physical event that has an explanation *in terms of regularities* has a sufficient explanation *in terms of physical regularities*. This version of the principle is compatible the explanatory knowledge argument, because it only implies that regularities have an explanation in terms of non-physical powers, not in terms of any non-physical regularities. But the principle is not compatible with interactionist dualism (except in its highly inelegant overdeterminist version), because interactionist dualism posits the existence of psychophysical regularities, that are irreducible to physical regularities, to explain physical events in the brain or body. But the principle is compatible with Russellian monism, both the general version which takes all phenomenal properties to be explanatory relevant in virtue of realizing physical structure (as opposed to by adding further structure in the form of psychophysical regularities), and the specific version which takes phenomenal properties to realize physical structure in virtue of constituting the powerful grounds of regularities. So, the explanatory knowledge argument is fully compatible with physical explanatory closure, because it is also compatible with Russellian monism.

In conclusion, I have defended the explanatory knowledge argument, according to which all physical facts are knowable without experience, but some explanatory facts are not knowable without experience—namely phenomenal facts that ultimately explain and non-inductively predict regularities such as "pain makes subjects try to avoid it *ceteris absentibus*". Therefore, some facts are non-physical.

The explanatory argument resists the main objections to the original knowledge argument, the ability hypothesis and the phenomenal concept strategy. It is compatible with, and may have some dialectical advantages over, the previously proposed normative knowledge argument. It also suggests an explanatory role for phenomenal properties that is compatible with a plausible version of physical explanatory closure, in accordance with Russellian monism. It thereby confirms and deepens the challenge for a physicalist account of the phenomenal.[17]

[17] Many thanks to Sam Coleman, David Chalmers, Sebastian Watzl, Torfinn Huvenes, John Morrison, Insa Lawler and participants at the NorMind inaugural workshop (Bergen 2015), The Science of Consciousness (Helsinki 2015) and NYU Consciousness discussion group (New York 2015) for helpful comments and discussion.

Bibliography

Alter, Torin, and Yujin Nagasawa. 2012. What Is Russellian Monism? *Journal of Consciousness Studies* 19 (9-10): 67-95.

Armstrong, David M. 1978. *A Theory of Universals. Universals and Scientific Realism Vol. Ii*. Cambridge University Press.

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

Cartwright, Nancy. 1999. *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Chalmers, David J. 2003. Consciousness and Its Place in Nature. In *Blackwell Guide to Philosophy of Mind*, eds. S. P. Stich and T. A. Warfield. Malden, MA: Blackwell.

Chalmers, David J. 2010a. *The Character of Consciousness*. New York: Oxford University Press.

Chalmers, David J. 2010b. The Content of Phenomenal Concepts. In *The Character of Consciousness*. New York: Oxford University Press.

Chalmers, David J. 2013. Panpsychism and Panprotopsychism. *The Amherst Lecture in Philosophy* 8 (1-35.): Reprinted in Brüntrup and Jaskolla 2016.

Cutter, B., and M. Tye. 2014. Pains and Reasons: Why It Is Rational to Kill the Messenger. *Philosophical Quarterly* 64 (256): 423-433.

Davidson, Donald. 1970/1980. How Is Weakness of the Will Possible? In *Essays on Actions and Events*. Oxford: Clarendon Press.

Dretske, Fred I. 1977. Laws of Nature. *Philosophy of Science* 44 (2): 248-268.

Ellis, Brian. 2010. An Essentialist Perspective on the Problem of Induction. *Principia* 2 (1): 103-124.

Grahek, Nikola. 2007. *Feeling Pain and Being in Pain*. Cambridge, MA: MIT Press.

Hume, David. 1739. *A Treatise of Human Nature*.

Jackson, Frank. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32 (April): 127-136.

Jackson, Frank. 1986. What Mary Didn't Know. *The Journal of Philosophy* 83 (5): 291-295.

Kahane, Guy. 2010. Feeling Pain for the Very First Time: The Normative Knowledge Argument. *Philosophy and Phenomenological Research* 80 (1): 20-49.

Lewis, David. 1973. Causation. *Journal of Philosophy* 70 (17): 556-567.

Lewis, David. 1983. Postscript to 'Mad Pain and Martian Pain'. In *Philosophical Papers*. Oxford: Oxford University Press.

Loar, Brian. 1997. Phenomenal States Ii. In *The Nature of Consciousness: Philosophical Debates*, eds. N. Block, O. Flanagan and G. Güzeldere The Mit Press.

Michotte, Albert. 1963. The Perception of Causality.

Mumford, Stephen. 2004. *Laws in Nature*. New York: Routledge.

Nemirow, Laurence. 1979. *Functionalism and the Subjective Quality of Experience*.

Papineau, David. 2001. The Rise of Physicalism. In *Physicalism and Its Discontents*, eds. C. Gillett and B. Loewer. Cambridge: Cambridge University Press.

Papineau, David. 2002. *Thinking About Consciousness*. Oxford: Clarendon Press.

Papineau, David. 2006. Phenomenal and Perceptual Concepts. In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, eds. T. Alter and S. Walter Oxford University Press.

Perry, John. 1979. The Problem of the Essential Indexical. *Noûs* 13 (December): 3-21.

Perry, John. 2001. *Knowledge, Possibility, and Consciousness*. Vol. 301. Mit Press.

Shoemaker, Sydney. 1980. Causality and Properties. In *Time and Cause: Essays Presented to Richard Taylor*, ed. P. van Inwagen. Dordrecht: Reidel.

Spelke, Elizabeth S, Ann Phillips, and Amanda L Woodward. 1995. Infants' Knowledge of Object Motion and Human Action.

Stoljar, Daniel. 2001. Two Conceptions of the Physical. *Philosophy and Phenomenological Research* 62 (2): 253-281.

Strawson, Galen. 1987. Realism and Causation. *The Philosophical Quarterly* 37 (148): 253-277.

Strawson, Galen. 1994. *Mental Reality*. Cambridge, MA: MIT Press.

Strawson, Galen. 2008. Mental Ballistics: The Involuntariness of Spontaneity. In *Real Materialism*. Oxford: Clarendon Press.

Tooley, Michael. 1977. The Nature of Laws. *Canadian Journal of Philosophy* 7 (4): 667-98.

Wilson, Jessica M. 2006. On Characterizing the Physical. *Philosophical Studies* 131 (1): 61-99.