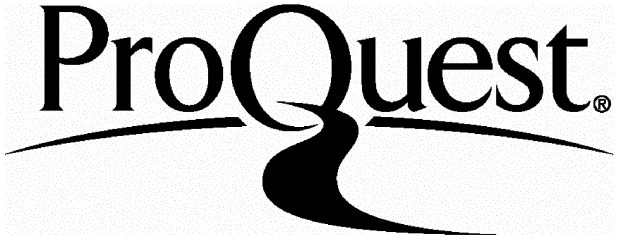
A Dissolution of the Problem of the Explanatory Gap

Jessica Alden Pepp
University College London
Submitted for the degree of MPhil in Philosophy



Abstract

This dissertation argues that the so-called 'explanatory gap' does not present a problem for physicalism. Following Frank Jackson, I define physicalism as the thesis that any possible world which is a minimal physical duplicate of the actual world is a duplicate simpliciter of the actual world. I define the explanatory gap as the non-deducibility of facts about conscious experience (such as the fact that seeing red is like this) from physical facts.

I argue that the existence of the explanatory gap amounts to the non-deducibility of a certain kind of fact – the kind of fact which one can only know if one has a particular experience – from facts that can be known without any experience in particular, such as facts given in terms of physics. I then argue that since this non-deducibility is a gap between different kinds of knowledge (that which we can have without particular experience, and that which requires particular experience to have) and not a gap between different kinds of objects or properties, it does not, prima facie, present a problem for physicalism. For physicalism, as defined, is a thesis about the objects and properties of the actual world, and is not a thesis about knowledge.

I next address arguments from Jackson and David Chalmers that if there is an explanatory gap, despite the prima facie compatibility of physicalism with the explanatory gap, then physicalism is false. I reply that their arguments for this conclusion depend on the use of the two-dimensional semantic framework, which cannot account for the kind of facts that require their knower to have a particular kind of experience. Thus, arguments such as Jackson's and Chalmers's, which use the two-dimensional framework to show that physicalism requires the deducibility of these kinds of facts from physical facts, cannot succeed.

Table of Contents

Chapter One: Introduction

The aim of this dissertation is to argue that the so-called 'explanatory gap' (so named by Joseph Levine, 1983, p. 357) is not a problem for physicalism. The argument I will present is compatibilist about physicalism and the explanatory gap. The existence of the explanatory gap, I will claim, is fully compatible with the truth of physicalism. Thus, anti-physicalists should not point to the existence of the explanatory gap as a way to attack physicalism, nor should physicalists feel the need to deny its existence in order to uphold physicalism.

The first two chapters of the dissertation will be devoted to the explication of the problem of the explanatory gap for physicalism. Before I can argue that physicalism is not threatened by the explanatory gap, I must make clear what physicalism is. Thus, in Chapter 2, I develop a clear, minimal notion of 'physicalism', which includes what I think all physicalists can accept as a necessary, if not sufficient, condition for the truth of physicalism. This is Frank Jackson's definition of physicalism as the thesis that any world which is a minimal physical duplicate of the actual world is a duplicate simpliciter of the actual world (1998, p. 12). Any possible world that is physically identical to the actual world is completely identical to the actual world. This condition, while necessary for all versions of physicalism, is not sufficient for some. Type physicalism, for example, holds not only that all properties are determined by, or 'supervenient' on, physical properties (which is all that is required by the Jacksonian version just presented) but also that all properties are *identical* to physical properties. I choose to work with the weaker version of physicalism because those who argue that the explanatory gap is a problem for physicalism take it to be a problem for *all* versions of physicalism. To establish this claim, they must demonstrate that it is a problem for even the most minimal version of physicalism.

In Chapter 3, I elucidate the explanatory gap and explain two ways in which it is alleged to present a problem for physicalism. I propose that the explanatory gap amounts to the claim that facts of conscious experience (such as the fact that seeing red is *like this*) cannot be deduced from physical facts (such as the fact that apples reflect light of a certain wavelength, the fact that light of that wavelength has a certain

effect on human retinas, and so on). When the explanatory gap is understood in this way, I argue, several different arguments establish that such a gap exists.

There are two ways in which the existence of the explanatory gap is taken to be a problem for physicalism. Some philosophers, such as Levine, Thomas Nagel, and Colin McGinn, argue that the explanatory gap presents an epistemological problem for physicalism. On this 'epistemological version' of the explanatory gap problem, the existence of the gap does not demonstrate the falsity of physicalism, but it does make the truth of physicalism mysterious, or paradoxical. On this version of the problem, we know that physicalism is true, but we do not understand how it could be in light of our inability to deduce facts about conscious experience from physical facts. The epistemological version of the explanatory gap problem will not be my central focus, since it is difficult to pin down in what sense it thinks the explanatory gap is a problem for physicalism. On this view, the explanatory gap does not imply that physicalism is false, yet it implies that physicalism is somehow lacking as a scientific theory. I will advance some considerations to suggest that the existence of the explanatory gap does not mean that physicalism is inadequate as a scientific theory. However, the main objective of the dissertation is to show that physicalism is not threatened by the explanatory gap; that the explanatory gap does not suggest it is false. I think supporters of the epistemological version of the explanatory gap problem generally agree with me on this point.

Thus, my argument will focus on the metaphysical version of the explanatory gap problem. According to this version of the problem, if there is an explanatory gap, then physicalism is false. This view of the problem is shared, for example, by anti-physicalist David Chalmers and Jackson, who is a physicalist. Both argue that the truth of physicalism requires it to be possible to deduce facts about conscious experience from physical facts. Chalmers thinks such a deduction is impossible; Jackson thinks it is possible. On this point, I will agree with Chalmers, arguing in Chapters 3 and 4 that facts about conscious experience cannot be deduced from physical facts. There is an explanatory gap. However, my central objective will be to show that the conditional 'if there is an explanatory gap, then physicalism is false,' which both Jackson and Chalmers espouse, is false.

To make this argument, I develop in Chapter 4 a set of distinctions between different notions of 'fact'. Since I want to show that for physicalism to be true it is not required that all facts (including facts about conscious experience) be deducible from physical facts, I first need to say what the word 'facts' means. Through analysis of Jackson's 'knowledge argument' (1982), I introduce a two-level distinction among different understandings of what a 'fact' is. At the first level, I divide facts into objects of knowledge (what I call '$facts_2$'), on the one hand, and the collections of objects and properties that make these $facts_2$ true (what I call '$facts_1$'), on the other hand. At the second level, I distinguish between two kinds of $facts_2$: those that can be known without their knower having any particular experience (what I call $facts_{2.1}$), and those which require one to have a particular experience in order to know them (what I call $facts_{2.2}$).

I argue that the knowledge argument demonstrates that there are $facts_{2.2}$: facts about conscious experience, such as the fact that seeing red is like this, are such facts. These facts are not deducible from physical facts. They can only be known if their knower has the experience of seeing red. The existence of $facts_{2.2}$, I will argue, is what the existence of the explanatory gap amounts to. I will also argue that the reinterpretation of the knowledge argument in terms of the different kinds of facts makes it clear that the existence of the explanatory gap in no way implies the falsity of physicalism, at least not prima facie.

In Chapter 5, I address Jackson's and Chalmers's arguments that physicalism is nonetheless committed to this kind of deducibility, though not prima facie. Drawing on the distinctions among different kinds of facts developed in Chapter 4, I attempt to show that their respective arguments for this conclusion do not succeed. Further, I argue that the two-dimensional semantic framework within which both Chalmers and Jackson couch their arguments cannot account for $facts_{2.2}$. Thus, I suggest that the existence of $facts_{2.2}$ does not so much demonstrate the falsity of physicalism, as it does the inadequacy of the two-dimensional semantic framework for analysing facts about conscious experience. Finally, I present an alternative way to argue that physicalism is not committed to the deducibility of facts about conscious experience from physical facts: this is an argument of Ned Block and Robert Stalnaker's (1999), to the effect that there is no such deducibility of any kind of fact (including everyday

macrophysical facts such as that water is H_2O) from basic physical facts; thus we should not expect it in the case of facts about conscious experience.

By the end of Chapter 5, I hope to have shown that the explanatory gap is not a problem for physicalism. I will not have shown that there are not other problems for physicalism. For instance, I will not have addressed the zombie argument: that the conceivability of a minimal physical duplicate of the actual world which is not a duplicate simpliciter (i.e., because some facts of conscious experience are different at that world from what they are at the actual world) entails the possibility of such a world, which entails that physicalism, as I have defined it, is false[1]. I am not concerned here to rebut anti-physicalist arguments relying on the entailment from conceivability to possibility. That would be a separate project. Indeed, one aim of this dissertation is to focus the question of whether or not physicalism is true away from the explanatory gap, and toward arguments like the zombie argument.

[1] See page 28 for a discussion of the zombie argument.

Chapter Two: What is Physicalism?

In this chapter, I will present and defend the understanding of the thesis of physicalism under which I will argue, in the remainder of the dissertation, that the explanatory gap is not a problem for physicalism. I define physicalism as, roughly, the thesis that all objects and properties are physical, or are supervenient on physical objects and properties. I define physical objects as those objects which either are or are fully constituted by the same kind of fundamental objects that constitute the non-conscious parts of the world.[2] I define physical properties as those properties which are or are instantiated by the same properties that instantiate non-conscious properties, or are instantiated by the same objects that instantiate non-conscious properties. All of these definitions require explication and defense, and this chapter will attempt to provide both.

The chapter has three parts. First, I will give a detailed account of my definitions of physicalism and the physical. Second, I will present briefly what seems to be the strongest argument in favour of physicalism about the mind – the causal argument for physicalism. As this dissertation is not an argument for physicalism, but a defense of it against objections stemming from the existence of the explanatory gap, I will not fill too much space with discussion of the causal argument. I present it only to establish physicalism's general plausibility. Third, I will submit it to the reader that prima facie, the version of physicalism I have developed is an ontological, not an epistemological, thesis. I will introduce the phenomenon I call the 'epistemicization of physicalism', in which it is claimed that physicalism is committed to certain epistemological requirements – in particular, epistemological requirements for which the existence of the so-called explanatory gap would be problematic. I will suggest (what I argue in depth in Chapter 5) that the epistemicization of physicalism does not succeed.

---

[2] Assuming that there *are* some non-conscious parts of the world. Pan-consciousness, under this definition of 'physical', would entail that nothing is physical. That is not to say that I am defining 'physical' as non-mental. If physicalism is true, the mental *is* physical under my definition, since the same kind of fundamental objects – physical objects – constitute conscious objects as constitute non-conscious objects, and the same kind of fundamental properties and objects – physical properties and objects – instantiate conscious properties as instantiate non-conscious properties.

2.1 A Definition of Physicalism

The definition of physicalism that I am using has two parts. The first part tells us what physicalism claims about physical objects and properties: physicalism claims that *all* objects and properties are either identical to or supervenient upon physical objects and properties. The second part tells us what the *physical* is: physical objects are objects which either are or are fully constituted by the same kind of fundamental objects that constitute the non-conscious parts of the world; physical properties are properties which are or are instantiated by the same properties that instantiate non-conscious properties, or are instantiated by the same objects that instantiate non-conscious properties. In explicating the definition, I will take each part in turn.

2.1.1 The physicalist claim

I understand the physicalist claim in broadly the same way that Jackson does. Jackson presents it in the following form: 'Any world which is a *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world' (1998a, p. 12). For a world to be a minimal physical duplicate of our world is for it to be identical to the actual world in every physical respect – to contain all the physical objects that the actual world contains and to instantiate all the physical properties that the actual world instantiates – and to neither contain any object that is not one of those physical objects, nor instantiate any property that is not one of those physical properties. For a world to be a duplicate simpliciter of our world is for it to contain all the objects (whether physical or non-physical) that the actual world contains and to instantiate all the properties (whether physical or non-physical) that the actual world instantiates. Thus, Jackson's version of physicalism says that a world which is identical to the actual world in all its *physical* objects and properties is identical to the actual world in *all* its properties, full stop.

Jackson's expression of the physicalist claim is useful for my purposes because it is a claim that all physicalists, be they type identity theorists, or non-reductive supervenience physicalists, should endorse. Of course, type identity theorists would argue that Jackson's version, while necessary for physicalism, is not sufficient for it. Type identity theory says that all properties and states (including

mental properties and states) are identical to physical properties and states. This implies that in a possible world in which all the same physical properties are instantiated as are instantiated in the actual world (which would be the case in a minimal physical duplicate of the actual world), *all* the same properties are instantiated as are instantiated in the actual world. For, according to type identity theory, all properties instantiated in the actual world are identical to physical properties, and, by definition, all physical properties instantiated in the actual world are also instantiated in the minimal physical duplicate. And, if all the same properties are instantiated in a minimal physical duplicate of the actual world as are instantiated in the actual world, then the minimal physical duplicate must also contain all the same objects as the actual world contains. For, as I will argue below, the sense of 'properties' that is relevant in discussions about whether physicalism is true (i.e., whether all objects and properties are physical or supervenient on physical objects and properties) is David Lewis's notion of 'natural properties' – roughly, those properties which determine the resemblances between and causal powers of objects (1983b, p. 192). By Leibniz's law, two worlds containing objects having all the same causal powers and bearing all the same resemblances to one another could not differ at all in the objects they contain. Thus, Jackson's formulation of physicalism is implied by type identity theories.

However, Jackson's formulation makes a weaker claim than these theories. Identity theory is commonly taken to be defeated by the problem of multiple realizability. The argument is roughly as follows. It seems plausible that pain could be realized in some non-human organism by a physical process other than c-fibres firing (assuming that c-fibres firing is the process that realizes the causal role for pain in humans). Some other process might just as well fill the causal role for pain and even have the qualitative feel of pain in that organism. Thus, the property of being in pain is not identical to the property of having c-fibres firing.

An alternative to type identity theory which is not vulnerable to the multiple realizability argument is token identity theory combined with supervenience physicalism. This is the variety of physicalism presented in Jackson's version of the

thesis.[3] Token identity theories say that mental tokens - that is, particular mental objects or events - are identical to physical tokens, while mental types - properties, or states - may merely *supervene* on physical objects, events, properties, and states. Thus, token identity theory might say that human pain (or, more accurately, a particular human pain) is identical to a particular physical event, but that pain in general is not identical to any particular physical state, though it does supervene on physical objects, events, or states.

The meaning of 'supervenience' is roughly what is implied by Jackson's formulation: the mental objects and properties which constitute and are instantiated in the actual world supervene on the physical objects and properties which constitute and are instantiated in the actual world just if a possible world which is constituted by all the same physical objects, and in which are instantiated all the same physical properties, as the actual world, also contains all the same mental objects (indeed, all the same objects, full stop) and has instantiated in it all the same mental properties (indeed, all the same properties, full stop) as the actual world. Thus, supervenience is different from identity. Identity is the relation between A and B such that A *is* B. Supervenience of A on B, on the other hand, is the relation between A and B such that if B, then A. Of course, if A is B, then it is also true that if B, then A. So identity of A and B entails supervenience of A on B (as well as supervenience of B on A). But supervenience of A on B (or of B on A) does not entail the identity of A and B.

It would seem that supervenience physicalism requires something a bit more than the modal claim that it is necessary, if a world has all the same physical objects and properties as the actual world, that it has all the same objects and properties as the actual world, full stop. For a dualist could allow that this is necessary, while maintaining that mental objects and properties are distinct from physical ones, not reducible to the same fundamental physical properties, and not causally efficacious in physical systems (Crane 2001, p. 58). Terence Horgan suggests that the appropriate relation, which he labels 'superdupervenience', might not be available (1993, p. 581).

[3] However, Jackson elsewhere endorses type identity theory (1995, pp. 263-268).

The task of characterizing the relation of supervenience in such a way that it rules out dualism without collapsing into identity is a difficult one. In addition, I do not think it is at all clear that supervenience physicalism is called for on account of the multiple realization argument. Why is it so impossible to think that the feelings which play the causal role for pain in non-humans - say, an octopus's pain, or a Martian's pain - might actually be different feelings from the feeling that plays the causal role for pain in humans? Why could two qualitatively different feelings not play roughly similar causal roles in different kinds of beings? If 'pain' is taken to refer, not to the causal role for pain, but to *the way pain feels*, then Martians might not feel pain in the same sense that humans do. Martians might have some feeling which fills the same causal role in them as the feeling of pain fills in humans, but they might not actually have the feeling of pain in the sense that we mean 'pain'. We might refer to the Martian's pain and our pain by the same word just because they play roughly similar causal roles, but this need not be anything more than a way of talking. If what we really mean by pain is a particular qualitative feeling, and not the fulfillment of a certain causal role, then there is no reason to think that the same feeling must be present in all kinds of beings in which some feeling plays the causal role that pain plays in humans.

I will not adjudicate here the question of whether and how the notion of supervenience can be fully characterized, or whether supervenience physicalism holds much advantage over type physicalism. For my purposes in this dissertation it does not much matter. Supervenience physicalism is a weaker form of physicalism than type physicalism; thus it is harder to refute. This means that a refutation of supervenience physicalism by the problem of the explanatory gap would be a stronger refutation of physicalism than a refutation of type physicalism. Thus, I am more interested to deny that the explanatory gap can refute supervenience physicalism than I am to deny that it can refute type physicalism. For this methodological reason, then, I choose to construe the objection to physicalism from the explanatory gap as an objection to supervenience physicalism.

2.1.2 The meaning of 'physical'

The second part of my definition of physicalism is the definition of 'physical'. I need to explain what it means for an object or property to be 'physical'. This question famously threatens to pull us into 'Hempel's dilemma' (Levine 2001, p. 18). This dilemma is a choice between defining physical objects and properties as those objects and properties which are described by current physics, or as those objects and properties which will be described by a future, completed physics. If we choose the former definition, we make physicalism almost certain to be false, because physics adds to its catalogue of fundamental objects and properties, and we have no reason to think the catalogue is complete at present. Thus, it is not likely true that all objects and properties are or supervene on objects and properties that are described by current physics.

On the other hand, if we choose the latter definition, we say that physics includes whatever fundamental objects and properties we turn out to need for causal explanations of all phenomena. This could include fundamental phenomenal properties like the ones Chalmers advocates as the basis of conscious experience. Such properties would be instantiated in the actual world because of fundamental psycho-physical laws which are not themselves supervenient on physical laws (1996, p. 126). In a scenario such as Chalmers proposes, we would hardly want to say that these phenomenal properties, which are causally isolated from everything else physics describes except because of some similarly isolated non-physical laws, are physical properties. But if we define the physical as including whatever fundamental objects and properties we turn out to need to causally explain all phenomena, these properties would be included, since we need them to causally explain conscious phenomena. Thus, as Levine says, mental properties would be made physical 'by fiat'. This conception of the physical leaves out something crucial about what is meant by 'physical', which is that everything which is physical should cohere within a complete physical system. Physics should not include, to use Herbert Feigl's words, 'nomological danglers' (1958, p. 139): laws which connect physical objects and properties with mental objects and properties without there being any further connection between those mental objects and properties and other objects and properties described by physics.

I will adopt an interpretation of 'physical' which does not require specification of the fundamental objects and properties, currently known or to be known later, that are included in physics. On this interpretation, physics is whatever theory fundamentally describes everything that is uncontroversially non-conscious. Examples of the uncontroversially non-conscious are boxes, lamps, postcards, clouds, whiskey, and ultraviolet light. We do not know what this theory will *say* when it is complete, but we do know that it will fundamentally describe such entities. This is just what we mean by 'physics'. Given this understanding of 'physics', 'physical objects' and 'physical properties' can be understood as follows:

Physical objects are those objects over which the theory that fundamentally describes uncontroversially non-conscious objects (i.e., physics) quantifies.

Physical properties are the properties over which the laws of physics quantify.

This interpretation, which can be found in various forms in Levine (2001), Papineau (2002), and Fodor (1974), captures what we mean by 'physical' when we wonder whether conscious experience is physical. What we are wondering is whether there is something about conscious experience that is fundamentally different from everything else. Whatever fundamental objects and properties non-conscious things reduce to (or supervene on), do conscious things reduce to (or supervene on) the same fundamental objects and properties? If they do, then physicalism, as I have defined it, is true, because any minimal physical duplicate of the actual world will be a duplicate of the actual world simpliciter. If, as on Chalmers's view, conscious experience reduces to or supervenes on *different* fundamental objects and/or properties from those to which uncontroversially non-conscious entities reduce or on which they supervene, then physicalism is false. For in that case, a minimal physical duplicate of the actual world will not necessarily be a duplicate simpliciter of the actual world. A world in which all the physical objects and properties (as defined) are the same as in the actual world will not necessarily be a world in which the objects and properties on which conscious objects and properties supervene are the same as in the actual world.

One might object to this interpretation that it would class panpsychism as a version of physicalism. For, if irreducibly psychical objects and properties turn out to

be among the fundamental objects and properties needed to describe even uncontroversially non-conscious phenomena, these objects and properties would be defined as physical. Therefore, the fact that such irreducibly psychical objects and properties were required to describe conscious phenomena would not make conscious phenomena (if they, too, reduced to or supervened on irreducibly psychical objects and properties) non-physical. Furthermore, in this case a minimal physical duplicate of the actual world would be a duplicate simpliciter of the actual world, making physicalism true according to my understanding of physicalism. Yet it seems that the existence of irreducibly psychical objects and properties runs exactly counter to the spirit of physicalism, and should be excluded under the definition of the thesis.

In response to this I would suggest that if indeed irreducibly psychical objects and properties must be postulated to explain even those phenomena which I have classed as uncontroversially non-conscious, then it does not seem unreasonable to include such objects and properties in the subject matter of physics. For if panpsychism held, then these objects and properties must be included in a description of the uncontroversially non-conscious phenomena which physics, according to any understanding we could have of it, must describe if it describes anything at all.

Levine has a way out of the panpsychism problem: he argues that physicalism should be understood simply as the thesis that 'only non-mental properties are instantiated in a basic way; all mental properties are instantiated by being realized by the instantiation of other, non-mental properties' (2001, p. 21). This characterization of physicalism would exclude panpsychism: the instantiation of irreducibly psychical, or mental, properties, whether by conscious or uncontroversially non-conscious objects, would mean that physicalism is false. However, as Levine admits, this definition removes the privileged role of physics from physicalism. There need not be, on this version of physicalism, a unified theory – physics – that *describes* all the basic properties. On my version of physicalism, by contrast, the privileged role for physics is upheld.

Adjudication between Levine's version of physicalism and my version turns on whether it is more important for physicalism to rule out panpsychism or to privilege physics as the science out of which the theory describing all phenomena will

come. I am opting for the latter. Physicalism is the thesis that the actual world includes only a certain kind of entity – the kind treated of by physics, the science which aims to describe the fundamental nature of uncontroversially non-conscious things. It seems less important that physicalism rule out the possibility that this fundamental nature is mental, than that it maintain that it is *physics* which describes the entire actual world.

2.1.3 The meaning of 'objects' and 'properties'

As I use them, the terms 'objects' and 'properties' refer to the basic ontological building blocks of the world. Objects and properties are the fundamental categories into which all entities can be sorted.[4] Properties are features or qualities of objects, and objects are entities which instantiate those properties. Objects and properties constitute all the facts of the world. Although I am characterizing the basic entities in this way, nothing in my argument turns on this characterization. One might reject the division of the world into objects and properties, instead taking facts as the fundamental building blocks of the world. On this view, it is the facts of the world that exist, and the notions of objects and properties are abstractions from these facts. Someone who cuts the world along these lines can simply translate my talk of objects and properties into their view of what is fundamental. Nothing I will argue depends on any particular ontology.

Like Lewis, I will formulate physicalism as a thesis about the nature of objects and what Lewis calls 'natural properties'[5]. Lewis states the thesis of physicalism as follows:

> Among worlds where no natural properties alien to our world are instantiated, no two differ without differing physically; any two such worlds that are exactly alike physically are duplicates (1983b, p. 364).

---

[4] I would also add events as a basic category, on the grounds that objects differ from events in that objects exist completely at any time when they exist at all, whereas an event does not exist in its entirety at any one instant in which it is occurring at all. But by contrast, a four-dimensionalist view of objects would hold that objects themselves exist in four dimensions, and have temporal parts: thus, an event is just an object, not a different kind of entity. For the purposes of this dissertation, I can be neutral on this issue.

[5] I am not assuming Lewis's conception of properties as sets: it does not matter for my purposes whether properties are conceived of as sets or as universals.

Lewis defines 'natural properties' as 'the ones whose sharing makes for resemblance, and the ones relevant to causal powers' (p. 347). He also analyzes 'natural properties' as being intrinsic (p. 357). He contrasts 'natural properties' with 'abundant' properties, of which any class of things is one. Thus there are an infinite number of abundant properties instantiated in the actual world. Any two things, no matter how similar or dissimilar to one another, differ in countless abundant properties and have countless abundant properties in common. (p. 346). Thus, Lewis suggests that in analysing whether two things (i.e., two possible worlds) are duplicates, the relevant question to ask is whether they have all the same natural properties (p. 356). For no two things have all the same abundant properties.[6]

One can see that Lewis's formulation of physicalism is the same as Jackson's version, only with the relevance of natural properties made explicit. The first clause of Lewis's formulation explains what should be meant by 'minimal' as used in Jackson's version of physicalism, and the second clause simply says that any minimal physical duplicate of the actual world is a duplicate simpliciter. Lewis's formulation is slightly different from Jackson's in that it says that any two possible worlds at which no properties are instantiated that are not instantiated at the actual world, if they are physical duplicates, are duplicates simpliciter, rather than retaining the actual world as the comparison point, but the substance of the formulations is the same. Thus, the version of physicalism I want to discuss says that any world which contains all and only the physical objects that the actual world contains, and in which are instantiated all and only the natural physical properties that are instantiated in the actual world, contains *all* the objects that the actual world contains, and has instantiated in it *all* the natural properties that are instantiated in the actual world.

2.2 The Causal Argument for Physicalism

Having explained what I take the thesis of physicalism to be, I will now set forth briefly what David Papineau calls the 'canonical argument for materialism' (2002, p. 17): the causal argument. As I said in the introduction to this chapter, I will

[6] As on the question of whether objects and properties are the fundamental categories of existence, and on the question of whether objects have temporal parts, I can remain neutral on the question of whether

not evaluate the causal argument in any great depth. I present it here for two reasons. First, the causal argument establishes that physicalism is at least a plausible thesis. Thus, evaluating attacks against physicalism that are unrelated to the causal argument (such as the objections from the explanatory gap that I will consider) is a worthwhile project. Second, the causal argument for physicalism combined with the existence of the explanatory gap is what makes philosophers such as Levine, McGinn, and Nagel think physicalism is paradoxical: on the one hand, they say, the causal argument shows conclusively that physicalism is true; on the other hand, the existence of the explanatory gap makes us unable to understand how physicalism *could* be true[7]. Since I will argue that the explanatory gap does not entail the paradoxicality of physicalism, it will be important to understand the causal argument, which represents one side of the supposed paradox.

The causal argument for physicalism can be sketched as follows[8]:

1. Conscious objects and properties have physical effects.
2. All physical effects are fully caused by physical causes.
3. There is not (at least, not always) causal overdetermination of the physical effects of conscious causes.
4. Therefore, conscious objects and properties must be identical to physical objects and properties.[9]

---

properties are universals, as D. M. Armstrong holds (1978, p. 61) or sets, as Lewis holds. As Lewis says, Armstrong's universals would pick out what Lewis calls natural properties (p. 347).

[7] See Levine (2001), McGinn (1991, p. 17), and Nagel (1974, p. 525).

[8] This roughly follows the explication in Papineau (2002, pp. 17-18).

[9] As Papineau notes, if the argument is made in terms of objects or events instead of properties, it will not generate the conclusion that conscious properties are physical properties, but only the weaker conclusion that conscious objects are physical objects. However, in order to generate the stronger conclusion, the argument would simply have to be rephrased as follows:

1'. Conscious objects/events have physical effects in virtue of their conscious properties.
2'. All physical objects/events are fully caused by physical objects/events in virtue of their physical properties.
3'. There is not (at least, not always) causal overdetermination of the physical effects of conscious causes in virtue of two different kinds of properties.
4'. Therefore, conscious objects *and* properties must be identical to physical objects and properties.

I have presented the argument in terms of objects and properties, but, in keeping with the discussion of section 2.1.3, it could just as easily be stated in terms of facts.

Premise 1 expresses the intuition that our conscious states have physical effects: my feeling of pain causes me to cry out, the gnawing ache I feel in my stomach causes me to seek food, and so on. Premise 2 is the completeness of physics, a doctrine now well-established by science.[10] Premise 3 expresses the intuition, which is supported by scientific methodological considerations, that there is not overdetermination of the physical effects of conscious causes: it is not the case that if I had not *felt* hungry I would still have sought food simply because I had a particular brain activity, nor that if I had not had the brain activity, I still would have sought food because I felt hungry. Rather, these 'two' causes seem to be inextricably linked. It is hard to see how these causes could be distinct unless some mechanism were postulated whereby whenever the one cause causes the physical effect, the other causes it as well. But such mechanisms are not present elsewhere in science, Papineau argues, and scientific methodological considerations count against postulating them. (p. 23.)

From premises 1, 2, and 3, the conclusion, 4, follows: if conscious objects and properties are the causes of some physical effects, and yet the only causes of these physical effects (given that they are not overdetermined) are physical, then conscious objects and properties must be physical. Of course, each of the three premises can be disputed. Against premise 1, it can be argued that epiphenomenalism (the view that conscious objects and properties do not have physical effects) is a plausible description of the relation between conscious objects and properties and physical objects and properties. Against premise 2, it can be objected that the completeness of physics has not been proven by science, only strongly indicated. Against premise 3, it can be argued that there could be mechanisms ensuring that the distinct conscious and physical causes of physical effects always occur so as to overdetermine these effects. I will not address these objections here, as it is not my purpose to defend the causal argument.[11] Rather, I want to reject the notion that the explanatory gap is a problem for physicalism, while assuming that the causal argument has established at least the initial plausibility of physicalism.

---

[10] See Papineau (2002, pp. 232-256) for a history of the establishment of the completeness of physics.

[11] Papineau provides a clear summary of these objections and physicalist responses to them (2002, pp. 21-28).

The causal argument for physicalism, as presented so far, establishes the identity of conscious objects and properties with physical objects and properties. However, the version of physicalism that I want to defend does not require conscious properties to be identical with physical properties. It only requires that conscious properties supervene on physical properties, such that in any possible world in which all the same physical objects are present and all the same physical properties are instantiated as in the actual world, all of the same objects as in the actual world will be present and all the same properties as instantiated in the actual world will be instantiated.

The causal argument does not seem to support this version of physicalism, for on this version of physicalism conscious properties are not supposed to be identical to physical properties, but only supervenient on them. Yet it is the former claim that the causal argument demonstrates. To make the causal argument demonstrate the weaker, supervenience claim, we must interpret 'cause' such that a property which supervenes on some set of other properties 'causes' whatever that set of properties causes. So, in this sense of 'cause', a flood 'causes' whatever damage all the individual H_2O molecules moving in particular ways cause. Also in this sense of 'cause', a feeling of hunger 'causes' whatever bodily actions are caused by all the various neural activities on which that feeling supervenes. If we understand 'cause' in this way, then the argument can be rephrased as follows[12]:

1". Conscious objects and properties cause (at least in the supervenience sense just described) physical effects.

2". All physical effects are fully caused by physical causes.

3". There is not (at least, not always) causal overdetermination of the physical effects of conscious causes.

4". Therefore, conscious objects and properties must at least be supervenient on physical objects and properties.

If one accepts the original, identity version of the causal argument, whether one also accepts this supervenience version will depend on whether one accepts that supervenient properties can be understood as causing the effects that their

[12] Again, this restatement follows Papineau (2002, p. 32).

supervenience bases cause. I will not argue this point here, though I will make one suggestion: it seems plausible that the transference of causation from supervenience base to supervenient properties might be part of what is included in the notion of 'supervenience' beyond the notion of necessitation. Not only does the supervenience base necessitate the supervenient properties, but also whatever the supervenience base causes, the supervenient properties also cause. This might move the notion of supervenience a step closer to Horgan's 'superdupervenience'.

In any event, whether or not the causal argument directly supports supervenience physicalism is not terribly important for the rest of this dissertation. It directly supports the identification of conscious objects and properties with physical objects and properties. And this identification entails that conscious objects and properties supervene on physical ones. So the causal argument at least indirectly supports supervenience physicalism. Recalling the two reasons I gave for introducing the causal argument at the beginning of this section, it is clear that just the identity form of the causal argument is sufficient for both. Initial support for type physicalism (directly) and supervenience physicalism (indirectly) has been provided. This establishes the value of defending physicalism against the independent challenge from the explanatory gap. In addition, the causal argument provides the reason that physicalism, though we do not understand how it could be true, *must* be true according to the 'paradox' views of Levine *et al.*

2.3 The Epistemicization of Physicalism

In the rest of the dissertation, I will be arguing that the explanatory gap (which I will discuss in the next chapter) is not a problem for physicalism. Those who take the explanatory gap to be a problem for physicalism do so on the grounds that the existence of the explanatory gap means that physics cannot *explain* facts about our conscious experience, and thus cannot provide a successful theory of the nature of this experience. I will take up this issue in the later chapters.

What I want to note at this stage is that the definition of physicalism I have developed does not include any epistemological requirements. It says that physicalism

is true if and only if something is true of all possible worlds that are minimal physical duplicates of the actual world - namely, it is true if and only if they are duplicates, full stop, of the actual world. So, for instance, any world that is a minimal physical duplicate of the actual world includes the same conscious experiences as the actual world. This does not imply, at least not prima facie, that for physicalism to be true one must be able to *know* a priori, from the knowledge that possible world *X* is a minimal physical duplicate of the actual world, that *X* includes the same conscious experiences as the actual world.

Nonetheless, some philosophers (e.g. Jackson and Chalmers) argue that the physicalist claim as I have defined it is indeed committed to our being able to have this kind of a priori knowledge, even though the commitment is not immediately obvious from the statement of the thesis. Other philosophers (e.g. Levine, McGinn, and Nagel) argue that without this a priori knowledge, physicalism would not be false, but it would be paradoxical or mysterious. For, on this kind of view, the causal argument would establish the truth of physicalism, but the a priori non-deducibility of facts about conscious experience from physical facts would make it impossible for us to understand *how* physicalism could be true. On both kinds of view, then, the a priori non-deducibility of facts about conscious experience from physical facts is taken to be a problem for physicalism. And as I will argue in the next chapter, arguments that there is an explanatory gap demonstrate just such a non-deducibility. Thus, these two kinds of view take the explanatory gap, in different ways, to be a problem for physicalism.

In the remaining chapters, I will present the arguments for both ways of epistemicizing physicalism. On the basis of the framework for understanding the explanatory gap that I will develop, I will argue that neither succeeds.

Chapter Three: The Explanatory Gap

The problem of the explanatory gap is the problem that if physicalism is true, it seems that we cannot explain conscious experience. I could know a complete physical description of the world (including my brain), the argument goes, but this knowledge would not explain to me why it is that the immediate experience I am having right now has the particular character that it does.

My aim in this chapter is threefold. First, in section 3.1, I attempt to distill a precise statement of what the explanatory gap is and why it is supposed to be a problem for physicalism. Then in section 3.2, I examine arguments for the existence of the explanatory gap, noting that they establish the explanatory gap as defined in section 3.1. Finally, I divide the way that the explanatory gap is used in arguing against physicalism into two categories: the metaphysical argument, that the explanatory gap implies the falsity of physicalism, and the epistemological argument, that the explanatory gap implies that physicalism, though true, is either mysterious or paradoxical – we do not fully understand how it could be true.

3.1 What is the Explanatory Gap?

The explanatory gap is supposed to be a problem for physicalism. It is supposed to be the problem that no description in terms of physics can ever explain conscious experience. Some philosophers argue that the inability of physics to explain conscious experience implies that physicalism must be false. Others take it to imply not that physicalism is false, but that if physicalism is true, we cannot explain how or why it is true. This is bad for physicalism, because physicalism is a scientific theory, and science is not just supposed to say that something is the case, it is supposed to say how and why it is the case; science is supposed to *explain* why things are as they are, not just say that they are in fact as they are. (Levine 2001, p. 69). I will discuss these two views of the implications of the explanatory gap in section 3.3 below. In this section, my goal is to clarify what is meant by saying that physics cannot explain conscious experience.

In the sentence 'Physics cannot explain conscious experience', the meanings of the subject (physics), verb (explain), and object (conscious experience) all are in need of illumination. To understand the explanatory gap, we must understand what conscious experience is, what it would mean for something to explain conscious experience, and what physics is.

I will begin with 'conscious experience'. Defining conscious experience is notoriously difficult, and I will leave the definition intuitive. Conscious experience is the way the world seems to us, our inner life, our awareness. It is the presence of the world for us, where the world includes both external objects and processes, as well as our own mental processes. I can have a conscious experience of seeing a spider crawling toward me on the table, as well as a conscious experience of thinking through a mathematical problem. These experiences can be contrasted with non-conscious states I might be in - digesting or regulating my breathing, for instance - and with unconscious states I might be in - stepping higher so as not to trip over a bump while walking, for instance. My being in such states does not have any qualitative character for me. In Nagel's famous phrase, it is not *like anything* for me to regulate my breathing or step higher. (Of course, it is like something to have regulated breathing, and would certainly be like something not to have regulated breathing, but the activity of regulating one's breathing itself is not like anything. I am not aware of it; it has no qualitative character. Similarly, it would be like something if I tripped, and it is like something not to trip, but the activity of adjusting the step so as not to trip is not like anything for me.)

In what follows, I will use 'conscious experience' and 'phenomenal experience' interchangeably. 'Phenomenal properties' and 'qualia' will refer to the properties of conscious experience.

Eliminativists about consciousness argue that such words do not properly describe any objects or properties. Thus, they 'eliminate' consciousness by saying that there is not really any such thing. If one is an eliminativist about consciousness, the explanatory gap is obviously not a problem, because there is no such thing as consciousness to be explained. I am not going to discuss the merits of eliminativism, except to say that it seems to deny the existence of the most obvious thing there is:

our experience. Physicalism as I have defined it is compatible with eliminativism, but I will not use arguments for eliminativism to defend physicalism. My goal is to defend the more general physicalist thesis, that all objects and properties are physical, against the objections that the explanatory gap falsifies this thesis or makes it unintelligible. I want to argue that physicalism can be both true and intelligible even if qualia are real properties, so I will assume their existence.

The next part of the statement ('Physics cannot explain consciousness') to explore is the notion of explanation. If the claim is that physics cannot explain consciousness, what would it mean for something to explain consciousness? One way of defining explanation is as follows: fact (or set of facts) *A* explains fact (or set of facts) *B* just if *B* can be deduced from *A* plus general laws. This is Carl Hempel's deductive-nomological ('D-N') conception of explanation (Hempel 1962, p. 18). The D-N conception is advocated by Levine (2001, p. 76). Levine acknowledges that, in practice, no phenomenon can ever be 'literally *deduced*' from the explanation offered for it, but maintains that a complete explanans, as difficult as it would be actually to produce one, would in principle allow one to deduce the explanandum. The actual explanations we give for phenomena are partial explanations that stand in for the complete ones. (p. 73).

Chalmers advocates a similar view of *reductive* explanation (he acknowledges that there are other kinds of useful explanation, such as historical explanation) (1996, p. 43). He characterizes reductive explanation as, roughly, the explanation of higher-level entities in terms of lower-level entities (p. 42). The way that one gets a reductive explanation of some phenomenon, Chalmers says, is by functionally analysing the phenomenon to be reduced, and then seeing which low-level facts fulfill the analysis (p. 51). Although Chalmers does not speak in terms of the D-N model, it is clear that his understanding of reductive explanation at least involves deduction: from knowledge of the low-level facts and one's analysis of the phenomenon to be explained (this analysis is conceptual, and therefore a priori), one can deduce the presence of the phenomenon. This amounts to the same deductive requirement Levine puts on explanation – that for there to be explanation, it is necessary that the explanandum be deducible from the explanans.

Although much has been written on the topic[13], I will not argue here about whether the D-N model of explanation is a good one. Instead, I want to address on their own terms the arguments that use it to suggest that the explanatory gap is a problem for physicalism. When Levine and others talk about the explanatory gap as the non-deducibility of facts about conscious experience from physical facts and laws, there is a certain kind of explanatory failure that they are pointing out. It is the same failure that the knowledge argument (which will be discussed briefly in this chapter and in detail in Chapter 4) uses to argue that physicalism is false.

This failure can be described as follows. One way for us to bring facts of conscious experience squarely within science would be to deduce them from the basic facts and laws of physics. If the facts of conscious experience could be shown to follow deductively from physical facts and laws in this way, then beyond explaining those physical facts and laws themselves, there would be nothing else to explain about the facts of conscious experience. The problem of explaining consciousness would then be comparable to the problems of explaining any physical phenomenon: it would be a purely empirical puzzle. But, as we will see, the knowledge argument demonstrates that we cannot deduce facts about conscious experience from physical facts and laws.[14] Thus, at least in this particular way, the failure of D-N explanation in the case of conscious experience sets the facts of conscious experience apart from physical facts: it creates a gap of a particular kind between them.

It is this gap – the explanatory gap – whose implications for physicalism I want to address. Thus, for the sake of argument, I will accept the D-N conception of explanation as the one that is involved in the explanatory gap. This is not a terribly far-fetched assumption to make. Developing a full and robust theory of explanation is a difficult task, and not one that has been satisfactorily dispatched. It is true that there are counterexamples to Hempel's D-N conception, but nonetheless it remains among the most fully worked out theories of explanation. Levine, Chalmers, and others argue

[13] A useful collection is Ruben 1993.

[14] Block and Stalnaker have argued that we cannot deduce any facts involving higher-order concepts, including standard scientific facts like the fact that water is H_2O, from basic physical facts and laws (2001). I will address their arguments in Chapter 5. For now, however, I want to accept the Levine/Chalmers line that while facts such as that water is H_2O are deducible from physical facts, facts such as that red looks like this are not.

that D-N explanation is a type of explanation which is problematically impossible under physicalism. I will argue that even if we accept D-N explanation as a meaningful understanding of explanation, its unavailability in trying to explain the facts about conscious experience by the physical facts is not a problem for physicalism.

The final step in explicating the statement of the explanatory gap is to say what 'physics' means. What is this body of information from which the facts about consciousness are not supposed to be deducible? Here I refer the reader to the definition, developed in Chapter 1, according to which physics is the theory that fundamentally describes everything that is uncontroversially non-conscious. Now we can restate the notion of the explanatory gap in the following way: from knowledge of the theory that fundamentally describes all uncontroversially non-conscious objects and properties, we would not be able to deduce all the facts about conscious experience.

3.2 Arguments for the Existence of the Explanatory Gap

The existence of this explanatory gap can be established in a few different ways. One way is through the thought experiment of Jackson's 'knowledge argument'[15]. Jackson's famous story of the colour-deprived physicist Mary probably does not bear repeating in its entirety, but here is a summary of the argument:

1. The physicist Mary knows all the physical information there is to know about colour vision while shut up in a black and white room.
2. Nonetheless, when Mary gets out of the black and white room and sees colour for the first time, she obviously learns something new; she gains a new piece of information.

[15] Howard Robinson made essentially the same argument in the same year (1982, p. 4). The spirit of the argument is also present in Nagel's 'What is it like to be a bat?' (1974). Further, the argument is prefigured by C. D. Broad's discussion of a mathematical 'archangel' who, despite understanding the basic structure of everything, cannot predict how things will smell (1925, p. 71) and by Russell's argument that a blind man could know all of physics and yet not know everything that a man who can see can know; thus 'the knowledge which other men have and [the blind man] has not is not a part of physics' (1927, p. 389).

3. Thus, Mary did not have all the information there is when she was stuck in the black and white room, so not all information is physical information. (1982, p. 471).

The complete knowledge argument has an additional premise – that according to physicalism, all information is physical – and thus reaches the conclusion that physicalism is false. I will discuss Jackson's conclusion in Chapter 4. For now, though, I want to focus just on his argument for 3: the conclusion that knowing all the physical information does not entail knowing all the information, full stop. It is this part of Jackson's argument that provides an argument for the existence of the explanatory gap. For if one could deduce facts about conscious experience from all the physical facts, then knowing all the physical information *would* entail knowing all the information (at least, all the information about conscious experience).

The argument from 1 to 3 is valid: if Mary knows all the physical information in the black and white room, and if she learns new information when she leaves, then all information must not be physical. Premise 1 is true by assumption. In Chapter 4, where I will discuss the knowledge argument in greater detail, I will address ways in which the truth of premise 2 can be questioned. I will also examine what the argument means by 'physical information' or 'physical facts', and I will explain why I think its epistemic conclusion, that from all the physical facts one cannot deduce all the facts there are, is correct. Ultimately, I will use this conclusion to argue that the explanatory gap is not really a problem for physicalism after all. At this point, however, I only want to note that the knowledge argument, if it is sound, demonstrates that there is an explanatory gap, in the sense that I defined 'explanatory gap' at the end of section 3.1. The knowledge argument shows that just by knowing all the facts of physics, Mary cannot deduce all the facts about conscious experience.

A second kind of argument that supports the existence of an explanatory gap is the 'zombie argument' against physicalism. The spirit of this kind of argument originates at least as far back as Descartes.[16] Descartes argued that since he could

[16] The use of the word 'zombie' to describe this kind of argument seems to originate with Kirk 1974, and is prefigured by Campbell's 1970 'imitation man'. William James also advanced a similar notion, with his 'automatic sweetheart' (Putnam 1999, p. 73).

imagine that his mind might exist while his body did not, it must be possible that his mind could exist while his body did not. Hence his mind and his body could not be the same thing, since one could exist without the other, but no one identical thing can exist without itself (1984, p. 54). The 'zombie' version of the argument simply reverses the thought experiment, asking us to imagine that our bodies exist while our minds do not, or at least while various aspects of our minds are altered or missing. In particular, the zombie argument appeals to our intuition that we can conceive of the existence of creatures – zombies – that are physically, structurally, and functionally identical to us humans, yet lack conscious experience.

The zombie argument can be sketched as follows:

A. Zombies are conceivable. (The 'zombie intuition'.)

B. The conceivability of zombies entails that zombies are possible. (The 'conceivability-entails-possibility' premise.)

C. If zombies are possible, then a minimal physical duplicate of the actual world is not necessarily a duplicate simpliciter.

D. Thus, physicalism is false.

The part of this argument that supports the existence of the explanatory gap is A, the zombie intuition. If it is conceivable that a creature might be exactly like us in every way that could be described by physics, and yet not be conscious at all, then we can conclude that from the physical facts alone we cannot deduce the facts about conscious experience. For if we could, then just by knowing that a creature (or more accurately, a world) is physically identical to us (to our world), we would know that that creature (world) is identical to us (to our world) in all facts about conscious experience as well. But the conceivability of zombies shows we do not know this. Thus, the existence of an explanatory gap as defined is supported by the zombie intuition.

I will not address the question of whether the zombie argument is a sound one against physicalism. I only mention the zombie argument to note that its first premise grounds the existence of the explanatory gap. My purpose is to argue that the explanatory gap *in itself* is not a problem for physicalism, and as such the zombie

argument is not my target. For the zombie argument does not say that the explanatory gap (i.e., the zombie intuition) is a problem for physicalism in itself. Rather, it says that the zombie intuition is a problem for physicalism *when combined with the conceivability-entails-possibility premise (B)*. There is a large literature devoted to establishing the truth or falsehood of this premise, but to survey it here would take me afield of my objectives. I would hope that one effect of my argument might be to focus antiphysicalist and physicalist attention alike away from the explanatory gap and onto arguments, such as the zombie argument, that proceed on independent or additional grounds (such as the conceivability-entails-possibility premise) toward the conclusion that physicalism is false.

Thus, both the knowledge argument and the zombie argument support the conclusion that there is an explanatory gap, as defined: we cannot deduce all the facts about conscious experience from the physical facts. If explanation of conscious experience by physics requires the deducibility of facts about conscious experience from physical facts, then the intermediate conclusion of the knowledge argument and the zombie intuition show that physics cannot explain conscious experience.

There are some additional arguments which can be used in their entirety or in parts to support the existence of the explanatory gap.[17] One is the argument from the inverted spectrum, which says that I can imagine a person physically, structurally, and functionally identical to me who nonetheless experiences red in the same circumstances in which I would experience blue, and for whom other colours are inverted in a similar way. The conceivability of this scenario demonstrates that knowing all the physical facts about me does not make it inconceivable that my conscious experience – of colour, in this case – might be different from what it actually is. (Chalmers 1996, p. 100.) This is essentially the same as the zombie intuition, except that it suggests that we cannot deduce the *nature* of experience from physical facts instead of that we cannot deduce the *existence* of experience from physical facts.

[17] I take these from Chalmers's summary of types of antiphysicalist arguments (1996, pp. 100-105).

Another argument is the argument from epistemic asymmetry, which is that we only know anything about consciousness, including that there is such a thing, from our own conscious experience. Thus, facts about consciousness cannot be deducible from any physical facts, which are available without first-person experience. (p. 102.) I will have more to say about this way of demonstrating the existence of an explanatory gap in Chapter 4.

Another way to demonstrate that there is an explanatory gap is through the 'absence of analysis' argument. This argument says that there is no analysis of the concepts of conscious experience such that a satisfaction of this analysis could be entailed by the physical facts. Functional and structural analyses do not capture what is meant by various 'phenomenal concepts', such as the way red looks or how a pinprick feel. Such concepts do not refer to the filling of causal roles or structural positions, but to experiences themselves, which have their own properties (qualia) independent of causal role or structural organization. (p. 105.) Thus, knowledge of physical facts could never allow one to deduce these phenomenal concepts.

Levine calls on an additional argument to demonstrate the existence of the explanatory gap. This is Saul Kripke's argument against the identity theory (Kripke, 1971). Kripke argues that if a mental state and a physical state are actually identical, they are necessarily identical. Thus, if the identity between a mental state and a physical state is contingent – i.e., if there is a possible world in which the mental state and the physical state are not identical – then the mental state and the physical state are not identical in the actual world, either. Kripke argues that mental-physical identities such as the identification of pain with the firing of c-fibres appear to be contingent in this way, because of the conceivability of scenarios like the zombie scenario described above.

Furthermore, Kripke argues, this appearance of contingency cannot be explained away as an illusion, as it can be, for example, in the case of the identity of lighting with an electrical discharge of ionized water molecules. In the lighting/electrical discharge case, it may *seem* conceivable to us that lightning might not have been this electrical discharge. However, Kripke says, what is really conceivable to us is a world in which something other than an electrical discharge –

and, thus, something other than lightning – fills the causal role by which we pick out lightning in the actual world. But this is not a possible world in which lightning is not an electrical discharge. Rather, it is a possible world in which the phenomenon which is exhibited as a jagged yellow flash across the sky, which splits trees and sets houses on fire, is not lightning.

'Lightning' is what Kripke calls a 'rigid designator'– it designates the same thing in all possible worlds – and that thing is an electrical discharge of ionized water molecules. 'Pain', he argues, is also a rigid designator. If pain is c-fibres firing in the actual world, pain is c-fibres firing in all possible worlds. Yet, as the zombie intuition demonstrates, this does not seem to be the case. And for the identity between pain and the firing of c-fibres, it is hard to see how the contingency suggested by the zombie intuition is an illusion. The causal role of lightning is something over and above lightning, which we use to pick out lightning. But the sensation of pain is not something over and above pain, which we use to pick out pain. Rather, pain just *is* the sensation of pain. So the possible world in which no pain is felt although c-fibres are firing is a possible world in which there *is* no pain despite there being c-fibres firing.

Kripke's argument furnishes an argument for the existence of the explanatory gap by highlighting the apparent contingency of the identity (supervenience) of conscious experience and (on) physical objects and properties. There seems to be a possible world in which conscious experience is not identical to (supervenient on) the physical. In the terms of the zombie argument, there is a conceivable world in which conscious experience is not identical to (supervenient on) the physical. If the facts about conscious experience were deducible from the facts of physics, pain would not seem to be only contingently identical to the firing of c-fibres, because the latter would logically entail the former.

Of course, Kripke takes his argument to imply that materialism is false. If an identity is true, he argues, it must be necessarily true, and yet for any psychophysical identity, there are possible worlds in which it is not true; hence no psychophysical identity can be true, and physicalism must be false. This argument, like the zombie argument, relies on the conceivability-entails-possibility premise. Kripke takes the conceivability of pain without c-fibres firing to entail the possibility of such a

scenario, on the grounds that we are not able to redescribe the pain without c-fibres firing scenario as one in which this is not really the case. As I have said, I will not enter the conceivability-entails-possibility debate. Kripke's demonstration of the apparent contingency in identities between concepts of conscious experience and concepts of physical properties is enough to demonstrate the existence of an explanatory gap. Levine calls such identities 'gappy' (2001, p. 81).

The arguments discussed in this section demonstrate that the facts about conscious experience are not deducible from physical facts: there is no conceptual necessary connection between physical facts and facts about conscious experience. In this sense, then, the arguments demonstrate that there is an explanatory gap.

3.3 The Metaphysical and Epistemological Versions of the Explanatory Gap Problem

But what follows from the conclusion that physics cannot 'explain' consciousness in this way? Proponents of the view that the explanatory gap is a problem for physicalism have answered this question in various ways, which can be divided into two general categories: the metaphysical version of the explanatory gap problem and the epistemological version of the problem.

The metaphysical version of the explanatory gap problem says that the explanatory gap's existence (which was established by the arguments in section 3.2) means that physicalism is false. So the metaphysical explanatory gap problem can be stated as follows:

I. One cannot deduce all the facts about conscious experience from physical facts. (There is an explanatory gap.)

II. Therefore, physicalism is false.

If there is to be a valid inference from I to II, another premise must be added in between them:

Ia. For physicalism to be true, one must be able (in principle) to deduce the facts about conscious experience from the physical facts.

Most of the remainder of this dissertation will be devoted to demonstrating that Ia is false. In Chapter 4 I will develop a framework for understanding the explanatory gap, teasing out just what it means for facts about conscious experience to be non-deducible from physical facts. Then, in Chapter 5, I will argue that given the understanding of the gap I will have developed in Chapter 4, premise Ia cannot be demonstrated in the ways that metaphysical gappists (I will focus on arguments from Jackson and Chalmers) argue it can be.

I will focus less on the epistemological version of the explanatory gap problem. On this version of the problem, the non-deducibility of facts about conscious experience from physical facts shows not that physicalism is false, but that it is inadequate or incomplete as a theory of mind because it does not explain the facts of conscious experience on the basis of physical facts. I take this kind of view, in varying forms, to be that of Levine (2001), McGinn (1991), and Nagel (1974). According to this view, there are strong independent reasons to believe that physicalism is true, such as the causal argument (see Chapter 2, Section 2.2). Further, epistemological gappists understand physicalism in the same way as I do: as a metaphysical thesis, without prima facie epistemological commitments. Thus, they do not take the non-deducibility of facts about conscious experience from physical facts to be sufficient to show that conscious experience is not identical to or supervenient on physical objects and properties.

However, epistemological gappists do take our inability to explain the facts of conscious experience just from the facts of physics (on the D-N model of explanation) to be a problem for physicalism. There are two ways in which epistemological gappism exhibits itself. In one way, as in the case of Levine's argument, physicalism is taken to be part of a paradox or antimony: the causal argument shows that physicalism is true, and yet, given the explanatory gap, we cannot understand how it could be true. In the other way, as in arguments such as Nagel's and McGinn's, physicalism is taken to be mysteriously true: the causal argument and considerations of simplicity show that physicalism must be true, but at the same time we cannot see how it is true, since we cannot deduce the facts of conscious experience from the physical facts. We know physicalism is true, but we cannot fully understand the physical-ness of conscious experience. This is either because it is beyond our current

conceptual abilities to understand – as I think is Nagel's view – or because the limits of the human mind place it forever outside our conceptual abilities – as McGinn argues.

The epistemological version of the explanatory gap can be stated as follows:

I. One cannot deduce all the facts about conscious experience from physical facts. (There is an explanatory gap.)

II'. Therefore, physicalism, though true, is mysterious or paradoxical, or in some other way inadequate or incomplete.

As did the metaphysical version of the explanatory gap problem, this inference needs an additional premise to be valid:

Ia'. For physicalism to be adequate as a theory of mind, the facts about conscious experience must be deducible from the facts of physics.

Ia' is a very different target from Ia of the metaphysical gap argument. Ia is a claim about a condition under which physicalism is false (i.e., physicalism is false if the facts about conscious experience are not deducible from physical facts). Thus, if one can show that physicalism is not necessarily false under that condition (as I will try to do in Chapter 5), then one has shown that Ia is false. But Ia' is a claim about a condition under which physicalism is inadequate. To refute it, one would have to show that physicalism is not necessarily inadequate under that condition.

Thus, the argument about whether the explanatory gap is an epistemological problem for physicalism will turn on what is required for physicalism to be adequate. Levine argues that for physicalism to be adequate, it must tell us not only *that* all objects and properties are identical to or supervenient on physical objects and properties (as the causal argument presumably does), but also why this is the case. 'Science is in the business of explanation,' Levine says (p. 69). And, at least on the D-N model of explanation which I have accepted for the sake of argument, complete explanation of facts about conscious experience by physical facts entails the in-principle deducibility of facts about conscious experience from physical facts (p. 76).

This question of what is required for physicalism to be adequate, beyond what is required for it to be true, is a difficult one to grasp. Metaphysical gappists like Jackson[18] and Chalmers are quite clear about the epistemological requirements they believe physicalism to have: if the facts about conscious experience are not in principle deducible from physical facts, then physicalism is false. But epistemological gappists are not clear about what the importance of such commitments is for physicalism. These epistemological commitments have some importance for physicalism, for if they are not fulfilled, physicalism is declared to be mysterious or paradoxical. But they do not have the status of *requirements* for the thesis, given that physicalism is declared true on independent grounds even if they are not fulfilled.

Since my aim is to argue that the explanatory gap is not a problem for physicalism, and since it is not clear what sort of problem for physicalism the epistemological version of the explanatory gap is supposed to be, I will not focus on it. But I would suggest that the 'business of explanation' science is in may not be D-N explanation of all facts by physical facts. I have accepted that the D-N model describes a certain kind of explanation which is lacking for the supervenience of facts about conscious experience on physical facts. Perhaps, if Jackson's and Chalmers's arguments are successful, the unavailability of this kind of explanation of facts about conscious experience by physical facts will be shown to entail the falsity of physicalism. But it is not clear, even if this kind of explanation turns out to be required for physicalism to be true, that such an explanation is either necessary or sufficient for physicalism to be explanatory in the way it ought to be *qua* scientific theory.

It would seem that scientific theories can be explanatory without involving a necessary deduction of the explanandum from a physical explanans. For instance, we might explain why a kitten has an extra claw on one foot by saying that it has a particular chromosomal mutation, without arguing that the presence of that mutation logically entails the presence of the extra claw. Levine might respond that if we knew

[18] Of course, Jackson is now a physicalist and does not believe the explanatory gap is a problem for physicalism. However, this is because he does not think there *is* an explanatory gap in the sense demonstrated by the knowledge argument: he now thinks that facts about conscious experience are in principle deducible from physical facts (Jackson 1998b). I classify Jackson as a metaphysical gappist because he subscribes to premise Ia: the view that if there is an explanatory gap, physicalism is false.

everything about the kitten's physiology in fundamental physical terms, the presence of the extra claw would indeed be logically entailed from this information plus the laws of nature. But we could respond to this that a complete description of the microphysical basis of the kitten's physiology would not actually explain why the kitten has the extra claw, just as giving someone a complete microphysical description of my brain would not explain to them why I am feeling sad, or happy, or unsettled. Biology, psychology, and the other 'special sciences' (roughly, the sciences other than physics), to use Jerry Fodor's term (1974), provide explanations for the phenomena they study in terms of the sciences of those phenomena. For an explanation in physical terms to be as informative as an explanation in the terms of the relevant special science, we would need bridge laws between the terms of physics and the terms of the special sciences.

As Fodor argues, there is no prima facie reason to think that such bridge laws are available. It does not seem that predicates which are explanatory in a special science like economics (Fodor uses the example of 'is a monetary system' (1974, p. 143)) have corresponding predicates that are explanatory in other sciences – in physics, for example. Contrary to this, Jackson argues that bridge laws can be provided in the form of conceptual analyses of higher-level terms: by analysing the concept of 'water', for instance, we can tell that H_2O (or its physical basis) is water (1998b, p. 28). In Chapter 5, I present Block and Stalnaker's (1999) argument that conceptual analyses do not provide bridge laws allowing the deduction of higher-level facts (facts about water, e.g.) from physical facts. But to address in depth the question of whether bridge laws between the sciences are available would take me beyond my present aims. My purpose here is only to raise doubts about the monolithic conception of explanation as deduction that the epistemological gappists espouse. The burden of proof that the bridge laws required for this kind of explanation are available lies with them, and it is a heavy burden.

In sum, I agree with the epistemological gappists that the explanatory gap represents *one* kind of explanation that physics cannot provide for consciousness. Further, I take seriously, and will address, the arguments of metaphysical gappists that the failure of this particular kind of explanation entails that physicalism is false. However, I do not see any compelling argument that such a failure, while not

demonstrating the falsity of physicalism, nonetheless demonstrates that physicalism has not dispatched its duties as a scientific theory. For scientific theories can be explanatory in different ways, and the fact that physicalism is not explanatory in one way does not mean it is not explanatory in other, more relevant ways. Thus, physicalism need not be any less adequate as a scientific theory of consciousness than the theory of the cat's mutated chromosome is as a scientific theory of the extra claw.

In the next chapter, I will present a framework within which to understand the explanatory gap. This framework will help me to demonstrate, in Chapter 5, that the explanatory gap does not falsify physicalism. Further, this framework will suggest that D-N explanation of facts about conscious experience by physical facts is not a requirement that should be placed on a scientific theory of consciousness, since facts about conscious experience are not the kinds of fact for which deduction is a relevant method of explanation.

Chapter Four: Keeping the Facts Straight

In this chapter, I will develop my argument that the explanatory gap, as characterized in Chapter 3, does not pose an epistemological problem for the thesis of physicalism, as characterized in Chapter 2. My discussion has two parts. First, I will examine one of the arguments that establishes the existence of the explanatory gap, the knowledge argument. I argue that the knowledge argument demonstrates the existence of a class of facts: those facts that cannot be known without the knower having had a particular experience. Second, I will argue that the existence of facts that require particular experience to know does not in itself suggest the falsity of physicalism, nor does it suggest that physicalism is epistemologically inadequate. That the existence of such facts does not entail the falsity of physicalism will be argued in greater depth in Chapter 5, where I defend this claim against the arguments of Chalmers and Jackson that physicalism is indeed committed to the deducibility of all facts from physical facts.

4.1 Interpreting the knowledge argument's conclusion

In Chapter 3, I discussed several arguments that can be used to demonstrate the existence of the explanatory gap, including the knowledge argument and the zombie argument. I argued that these arguments do establish that there is an explanatory gap, in the sense in which I have defined it: they show that from all the facts about the types of objects and properties that are fundamental to uncontroversially non-conscious entities (or 'physical facts'), it is not possible to deduce all the facts about conscious experience.

In this section, I will present one way to make sense of the knowledge argument's demonstration of the explanatory gap. The knowledge argument's conclusion (not its ultimate conclusion, that physicalism is false, but the intermediate conclusion that we cannot deduce all the facts about conscious experience from the physical facts) can be interpreted as a distinction between different types of facts. Interpreting the intermediate conclusion in this way provides a framework for assessing whether and how the explanatory gap poses a problem for physicalism.

To develop this interpretation, I will explicate the knowledge argument in somewhat greater detail than was provided in Chapter 3. Here is a paraphrase of the knowledge argument as Jackson originally presented it (1982, pp. 470-472):

1. The physicist Mary knows all the physical information or facts there are about colour vision while stuck in a black and white room. (Assumption.)
2. When Mary gets out of the black and white room and sees red for the first time, surely she learns new information or a new fact; she gains a new piece of knowledge. (Intuitive lesson of the story.)
3. Thus, Mary did not know all the information or facts there are when she was stuck in the black and white room. (Follows from 2.)
4. Thus, not all information is physical (Follows from 1 and 3).
5. Physicalism is the theory that all information is physical. (Definition.)
6. Thus, physicalism is false.

I will set aside 5 and 6 for now. However, my discussion of 1 through 4 should block the ultimate anti-physicalist conclusion reached in 6 by its clarification of the meanings of 4 and 5. Two concepts appealed to in 1 through 4 are in need of explication. The first is the concept of a fact, piece of information, or piece of knowledge. The second is the concept of a fact's, piece of information's, or piece of knowledge's being physical.

In his original presentation of the knowledge argument, Jackson uses the words 'knowledge' and 'information' so as to imply that information is what one has knowledge of. Knowledge of information describes both what Mary has in the black and white room and what she acquires when she leaves the black and white room and sees red for the first time:

> It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous *knowledge* was incomplete. But she had *all* the physical *information. Ergo*, there is more to have than that. (1982, p. 471; emphasis added on 'knowledge' and 'information'.)

Jackson's argument can be paraphrased such that what Mary knows in the black and white room and what she learns when she leaves are called 'facts' instead of

'information'. But however we choose to label the sort of thing that Mary has knowledge of in the black and white room and the sort of thing that she gains knowledge of upon leaving the black and white room, it is clear that the knowledge argument considers both to be the same sort of thing. In other words, the knowledge of information or facts that she has in the black and white room and the knowledge of information or facts that she gains upon leaving the black and white room and seeing red are different tokens of a single type of thing. The knowledge argument needs this to be the case, because otherwise it will not be able to reach the intermediate conclusion 4 (that not all information is physical) which, when combined with premise 5 (that physicalism requires all information to be physical) allows the knowledge argument to reach its ultimate anti-physicalist conclusion, 6.

To see that this is so, consider the following reformulation of the thesis of physicalism as put forth in 5:

5'. Physicalism is the thesis that all members of the set X (which we could refer to as the set of 'facts' or 'information' or 'objects of knowledge') are members of the set Y (which we could refer to as the set of 'physical facts' or 'physical information' or 'physical objects of knowledge').

For the knowledge argument to demonstrate the falsity of this thesis, it must demonstrate that at least one member of X is not a member of Y. It claims to have done this, in 1 through 4, by demonstrating that some members of X (namely those facts, information, or objects of knowledge that Mary learns when she sees red for the first time) are not members of Y. Clearly, the argument has to assume that both what she knows in the black and white room and what she learns when she sees red are members of X. Consider the alternatives. If neither were members of X, then the argument would have nothing to say about 5'. If only what she knows in the black and white room were members of X, and not what she learns upon seeing red, then the argument fails to show that any members of X are not members of Y (since it is assumed in 1 that what she knows in the black and white room are members of Y). Finally, the argument assumes that what Mary knows in the black and white room exhausts Y. Thus, if only what she learns upon seeing red are considered members of X, then the facts – the set X – trivially, and rather incoherently, exclude the physical facts. Thus, however the set X is labeled, the knowledge argument must hold that both

what Mary knows in the black and white room and what she learns upon seeing red are members of X. Both must be the same type of entity.

Some have argued that they are not the same type of entity, and that the knowledge argument fails as a result. Lewis (1988, pp. 514-518) and Laurence Nemirow (1990, pp. 492-495) have argued that while what Mary knows in the black and white room are *propositions*, what Mary learns when she sees red for the first time is something different. What she learns is an *ability*: she learns how to recognize red, imagine red, and so on. I will leave the notion of a 'proposition' very open, understanding propositions merely as the objects of factual knowledge, attributed in the form 'X knows that P', where P is the proposition. I take no stand on the metaphysics of propositions.

According to this 'Ability Hypothesis' (Lewis 1988, p. 514), when Mary leaves the black and white room, she gains knowledge of how to do some things (such as recognize and imagine the colour red) but she does not gain any knowledge *that* anything is the case; she does not gain any propositional knowledge. Since the knowledge she gains upon seeing red is a different kind of thing from the knowledge she had in the black and white room, the knowledge argument has not achieved its aim of proving that there are any facts, or information – members of X, as I put it – that are not physical facts or information – members of Y. All it has shown is that there are some abilities which are not members of Y, and this is not an interesting conclusion since by definition all members of Y are propositions, not abilities.

In response to the Ability Hypothesis, it has been argued that 'ability knowledge', or 'knowledge how', can be reduced to propositional knowledge, or 'knowledge that' (see Williamson and Stanley (2001), Moore (1997),[19] Snowdon (2002)). Thus, even if what Mary learns is an ability, having that ability is nothing over and above knowing a set of propositions, so what Mary learns is really the same type of object of knowledge as the propositions one could learn from a physics book, and the knowledge argument goes through. It has also been argued (Crane 2003) that even if ability knowledge does not reduce to propositional knowledge, and even if

Mary does gain ability knowledge when she sees red for the first time, Mary also gains some propositional knowledge when she sees red for the first time: namely, she learns that seeing red is like this. This gain in propositional knowledge, even if it is accompanied by a gain in non-propositional ability knowledge, is enough for the knowledge argument to go through. The knowledge argument shows that there are members of X – the set of all knowable propositions – which are not members of Y – the set of all propositions that are part of physics. For even if Mary knew all the propositions of physics, she would not know that seeing red is like this.[20]

The way that the Ability Hypothesis is supposed to demonstrate that the knowledge argument is unsound is by showing that what Mary knows in the black and white room and what she learns upon seeing red are not the same type of thing. For the knowledge argument to be sound, these must be the same type of thing: namely, they must be propositions, since Mary's knowledge of physics is knowledge of propositions. So what the knowledge argument means by 'information' or 'facts' must be knowable true propositions. If they are not the same type of object of knowledge, then, as Lewis and Nemirow argue, the knowledge argument would not even get off the ground.

In his original statement of the knowledge argument, quoted above, Jackson seems to be making the implicit claim that all knowledge is propositional, since he uses 'knowledge' interchangeably with 'having information'. And, as I have just argued, he is using 'information' to mean knowable proposition. But the knowledge argument need not presuppose that all knowledge is propositional. The knowledge argument only requires that *some* knowledge is propositional, and that Mary both has in the black and white room and gains upon release *some* propositional knowledge.

---

[19] Moore distinguishes knowing how from ability knowledge: for instance, a prisoner might know how to escape without having the ability to do so.

[20] I am assuming that states of knowledge are individuated by their objects, so that if there is a new state of knowledge, as there is when Mary learns that seeing red is like this, there is a new object of knowledge – the proposition that seeing red its like this. Against this, Mark Kalderon argues that there could be a new state of knowledge without a new object of knowledge; that states of knowledge could be partly individuated by their mode of justification, or the way in which they are known, as well as by their objects (Unpublished paper, 2002).

The second thing in the knowledge argument that needs further explication is the question of what it means for propositions to be 'physical', as those that Mary knows in the black and white room are described as being in premise 1. From the context of the 'Mary' thought experiment, we can discern two criteria for a proposition's being 'physical' in the sense of premise 1. The first criterion is that the proposition's subject matter – what the proposition is about – consists in, or is fully constituted by, objects over which the theory that fundamentally describes uncontroversially non-conscious objects (i.e., physics) quantifies. So for Mary to know all the physical information, she would have to know all the information about everything that is or is fully constituted by the fundamental objects of the theory of the uncontroversially non-conscious.

The second criterion concerns what we might call the 'mode' of Mary's knowing the proposition. According to the Mary story, Mary learns all the physical information about colour vision by watching a black and white television (471). What is the significance of the television being black and white? It is that Mary can learn all the physical information about colour vision *without experiencing colour vision.* Thus, the second criterion of knowable propositions' being 'physical' is that they must be knowable without their knowers experiencing their subject matter.

This criterion is a difficult one to characterize precisely. To know many propositions, we need to have *some* experience of what they are about. To know that the sky is blue, we need to have the experience of someone telling us that the sky is blue, or of reading that the sky is blue, or of being told that something else is blue whose colour we can match to the colour we observe in something which we are told is the sky. But we might think that to know the proposition that the sky looks *like this* requires a direct experience of what the proposition is about: the way the sky looks. We might think that being told or reading a description of how the sky looks does not suffice for knowing how it looks: the only way to know how it looks is to have the experience of looking at it. If there are any facts of this kind (facts one cannot know without having an experience of their subject matter), then Mary does not know any such facts about colour vision while in the black and white room. Thus, the set of all physical facts must exclude facts of this type, if there are any such facts. For physical

facts or propositions can be known without their knower experiencing their subject matter.

Using these explications of the notion of 'facts' or 'information' and of the notion of a fact's being physical, we can restate the first four premises of the knowledge argument as follows:

1'. The physicist Mary knows all the propositions that can be known without experiencing the phenomenon of colour vision herself, about those aspects of the phenomenon of colour vision which are or are fully constituted by the objects over which the theory that fundamentally describes all uncontroversially non-conscious phenomena quantifies. (Assumption.)

2'. When Mary gets out of the black and white room and sees red for the first time, surely she learns a new proposition – for instance, she learns that seeing red is like this. (Intuitive lesson of the story.)

3'. Thus, Mary did not know all the knowable propositions there are when she was stuck in the black and white room. (Follows from 2'.)

4'. Thus, not all knowable propositions are propositions that are *both*:

   a. About those aspects of phenomena which are or are fully constituted by the objects over which the theory that fundamentally describes all uncontroversially non-conscious phenomena quantifies; and

   b. Knowable without the knower himself experiencing the phenomenon that the proposition is about. (Follows from 1' and 3'.)

Now we can see how the existence of the explanatory gap, as defined, is demonstrated by the knowledge argument. The explanatory gap is the non-deducibility of all facts or information from physical facts or information. The knowledge argument demonstrates that if all the facts were deducible from the physical facts – those that conform to a and b under 4' above – then Mary would be able to know all knowable propositions in the black and white room. She would just deduce them from the physical propositions that, ex hypothesi, she knows in the black and white room. But Mary does not know all the knowable propositions in the black and white room, as the knowledge argument demonstrates. Thus, all knowable

propositions are not deducible from physical propositions. Thus, there is an explanatory gap.

But the restatement of premises 1 through 4 also suggests a different way to interpret the explanatory gap problem, and a way to object to premise 5 of the knowledge argument, the definition of physicalism as the thesis that all information or facts are physical. Let me begin with the interpretation of the explanatory gap problem. So far, I have drawn out three distinctions that the knowledge argument makes among different kinds of facts. The above analysis shows that the knowledge argument makes an implicit distinction between two different types of information or facts. This is the distinction drawn by criterion b, above: the distinction between propositions that are knowable without their knower experiencing their subject matter, and propositions that can only be known if their knower experiences their subject matter. In addition, I have explored the distinction between propositional and non-propositional knowledge: this distinction is implicit in the knowledge argument inasmuch as the knowledge argument depends on both what Mary knows in the black and white room and what she learns upon seeing red for the first time being the former kind of knowledge. Finally, criterion b under 4' above makes a distinction between information about physical phenomena and information about non-physical phenomena.

But there is another distinction to be made here, which is crucial to understanding what the knowledge argument actually demonstrates. This is the distinction between facts, information, or propositions in the sense of what can be known, and facts, information, or propositions in the sense of what exists. I will refer to the latter, the truth-makers, as $facts_1$, and to the former, the knowable truths, as $facts_2$. $Facts_1$ are collections of objects and properties that make $facts_2$ – what can be known about these objects and properties – true. That there is a distinction between $facts_1$ and $facts_2$ can be demonstrated by noting that 'Hesperus is Phosphorus' is surely a different $fact_2$ from 'Hesperus is Hesperus', although the $facts_1$ that make the two propositions true are identical. When the ancients learned that Hesperus is Phosphorus, they came to know something – information, a $fact_2$ – that they had not known when they only knew that Hesperus is Hesperus. Thus, $facts_2$ cannot be the same as $facts_1$.

The notion that there is a one to one relationship between the objects of Mary's knowledge and the actual objects and properties in the world (in other words, that for any physical object or property there is just one piece of knowledge to be had, hence when Mary knows all the physical facts she knows everything about all physical objects and properties) is called into great question by the Hesperus/Phosphorus example. In that example, there is just one object, with two names. There is no difference in the $facts_1$ underlying the $facts_2$ that Hesperus is Hesperus and that Hesperus is Phosphorus. But if there were a one-one relationship between the objects of knowledge ($facts_2$) and the $facts_1$, then there would be a difference in $facts_1$ between the two different $facts_2$ that Hesperus is Hesperus and that Hesperus is Phosphorus.

Having made the $facts_1$/$facts_2$ distinction, it is clear that the knowledge argument is concerned with $facts_2$, since it relies on our intuitions about what Mary *knows* at different stages of the game. We can now label the further distinction within $facts_2$ discussed above: the distinction between $facts_2$ that can be known without the knower experiencing their subject matter, which I will call $facts_{2.1}$, and $facts_2$ that *cannot* be known without the knower experiencing their subject matter, which I will call $facts_{2.2}$. The above interpretation of the first four premises of the knowledge argument shows that 'physical facts' or 'physical information' refers to $facts_{2.1}$: those facts that can be known without the knower experiencing their subject matter, such as the facts Mary knew about colour vision in the black and white room. So what the knowledge argument shows is that by knowing all the facts that can be known without the knower experiencing their subject matter ($facts_{2.1}$), one does not know all the facts. In particular, it demonstrates that just by knowing all the $facts_{2.1}$, one does not know all the $facts_{2.2}$, such as the fact that seeing red is like this.

This does not seem a terribly interesting conclusion, given that $facts_{2.2}$, by definition, can only be known if the knower experiences their subject matter herself, and the knowledge argument stipulates that Mary does not experience seeing red while she is in the black and white room, and does experience it upon leaving. But there is an interesting conclusion to draw from these first four premises: the conclusion that there *are* $facts_{2.2}$. The first four premises demonstrate that there are

facts$_2$ which can only be known if the knower has a particular experience (an experience of the subject matter of the facts), because they show that one could know all the facts$_2$ that one can know without having an experience of their subject matter, and still not know all the facts$_2$: thus there are facts$_2$ that cannot be known without the knower having an experience of their subject matter. There are what I have called facts$_{2.2}$.

Now, premise 4' of the knowledge argument as interpreted above appears at first glance to do more than just establish that there are facts$_{2.2}$. For premise 4' asserts that there are propositions which do not fulfill *both* criteria: *a* - that they are about the physical (as defined) aspects of phenomena, and *b* - that they are knowable without the knower experiencing the subject matter of the proposition. Putting the issue this way makes it seem as if it matters whether or not the proposition that Mary learns upon seeing red for the first time is *about* objects which are, or are fully constituted by the objects over which the theory that fundamentally describes all uncontroversially non-conscious objects quantifies. It might seem this way because that is what is required by criterion *a*. However, it is obvious that the demonstration that there are propositions which do not fulfill criterion *b* is alone sufficient for the establishment of 4'. And in fact, this is all that the knowledge argument demonstrates. It does not demonstrate that the subject matter, or corresponding facts$_1$ – of the propositions – facts$_{2.2}$ – that Mary learns when she sees red for the first time are different from the facts$_1$ corresponding to the propositions – facts$_{2.1}$ – that she knows about colour vision in the black and white room.

This examination of how the knowledge argument demonstrates the explanatory gap provides a better understanding of what the existence of the explanatory gap amounts to. What it amounts to is that there are facts$_{2.2}$. There are facts, or knowable propositions, which cannot be known unless the knower has an experience of their subject matter. Some of these facts$_{2.2}$ are facts about conscious experience: thus we cannot deduce all the facts about conscious experience from the physical facts, which are the facts$_{2.1}$. The explanatory gap is a gap between knowledge of facts$_{2.1}$ and facts$_{2.2}$: these are different kinds of facts, and the latter are not deducible from the former. On the definition of explanation used in the statement of the explanatory gap, facts$_{2.1}$ cannot explain facts$_{2.2}$. For, on this understanding of

explanation, for $facts_{2.1}$ to explain $facts_{2.2}$ the $facts_{2.2}$ would have to be deducible from $facts_{2.1}$. And, by definition, $facts_{2.2}$ cannot be deduced from any other facts, but can only be known if the knower experiences their subject matter.

This is an important rephrasing of the problem, because it characterizes problems arising from the explanatory gap as problems about facts that can be *learned* in a certain way (by experiencing their subject matter) rather than simply as problems about facts that have a particular *subject matter* (conscious experience). This way of stating the explanatory gap emphasizes that the explanatory gap, as demonstrated, is in no way a gap between physical objects and properties and conscious objects and properties, but only a gap between one kind of knowledge and another.

We can also redescribe the zombie intuition in terms of these distinctions among types of facts:

The zombie intuition: I can imagine the physical $facts_{2.1}$ being the same as they actually are, while the $facts_{2.2}$ about conscious experience are different from what they actually are.

This intuition demonstrates that I cannot deduce the $facts_{2.2}$ about conscious experience from physical $facts_{2.1}$, which, again is what the explanatory gap amounts to.

It would be a simple exercise to redescribe in terms of $facts_1$, $facts_2$, and $facts_{2.1}$ the other arguments that establish the existence of the explanatory gap. However, since the discussion that follows will focus on the zombie intuition and the knowledge argument, I will not carry out that exercise here.

4.2 Is the explanatory gap a problem for physicalism?

Equipped with the interpretation of the explanatory gap in terms of the different types of facts, I now return to the central question of the dissertation. Does this non-deducibility of $facts_{2.2}$ from $facts_{2.1}$ present a problem for physicalism? Does it suggest that physicalism is false (the metaphysical version of the explanatory gap problem) or inadequate (the epistemological version)?

We can observe first that the existence of facts$_{2.2}$ does not show that there are non-physical objects or properties, since the truthmakers, or facts$_1$, that underlie facts$_{2.2}$ could well be the same as underlie physical facts$_{2.1}$. Nothing in the knowledge argument, zombie intuition, or other arguments that can be used to argue for the existence of an explanatory gap (which we now see consists in the existence of facts$_{2.2}$) shows that this is not the case. Thus, the only way to argue from the existence of the explanatory gap to the falsity of physicalism is via the thesis I labeled Ia in Chapter 3[21]. This is the thesis that for physicalism to be true, one must be able (in principle) to deduce the facts about conscious experience from the physical facts. We can now rephrase Ia as follows:

Ia*. For physicalism to be true, one must be able (in principle) to deduce facts$_{2.2}$ from physical facts$_{2.1}$.

I have argued that the knowledge argument can be interpreted as establishing that there are facts$_{2.2}$ which cannot be deduced from any other facts; further, I have equated this intermediate conclusion of the knowledge argument with the conclusion that there is an explanatory gap, as defined. Thus, if Ia* is true, the knowledge argument demonstrates, by establishing that there is an explanatory gap, that physicalism is false. In the next chapter, I will argue that Ia* is not true. Thus, the existence of facts$_{2.2}$, or the existence of the explanatory gap, does not entail that physicalism is false.

In Chapter 3, I explained that I would not focus on the epistemological version of the explanatory gap problem, on the grounds that it is not clear exactly what sort of problem for physicalism the epistemological version *is*. At the end of that chapter, I also said that the framework I have introduced in the present chapter would reinforce my suggestion that D-N explanation of facts about conscious experience by physical facts ought not to be required of a scientific theory of consciousness.

According to epistemological gappists, the causal argument establishes that physicalism is true. The problem, they say, is that we cannot deduce facts about

[21] See page 33.

conscious experience from physical facts, so although we may know *that* conscious experience is identical to or supervenient on the physical, we do not know *how* this is the case. One response to this worry about the case for identity theory, or type physicalism, is that it does not make sense to ask *how* Q and P could be identical once we know *that* they are identical. Asking what makes Q and P identical is like asking what makes the Eiffel tower be the Eiffel tower. There is nothing in virtue of which anything is itself. If we know that Q and P are identical, then that is all there is to know about their identity. Identities, as Papineau says, do not require explanations (1995, p. 264).

But this response does not apply to supervenience relations: we can quite reasonably ask how (or why) Q supervenes on P. What is it about P that necessitates Q? If P logically entailed Q, then we would know what it was about P that necessitated Q: it would be P's logical entailment of Q that necessitated Q. But if P does not logically entail Q, as is the case for the supervenience of facts about conscious experience (call them Q) on physical facts (call them P), then we do not know how P necessitates Q.

The distinctions among different types of facts that I have developed in this chapter do not fill in the missing explanation of how P necessitates Q by showing that in fact P does logically entail Q. They do not, in other words, provide a D-N explanation of Q in terms of P. But they do suggest an explanation of why any explanation of how P necessitates Q that *can* be found will *not* be a D-N explanation. This explanation is as follows. Q consists of $facts_{2.2}$, which are facts that cannot be deduced from any other fact, but can only be known if the knower has a particular experience. The knowledge argument establishes this via a thought experiment, but there is no reason to think (at least in principle) that it could not be established empirically. If it were known that certain facts can only be known if their knowers have certain experiences, this would explain why, although physicalism is true, we cannot deduce facts about conscious experience from physical facts. Now we would seem to have a scientific theory with adequate explanatory power. The causal argument shows that our conscious experiences must be identical with or supervenient on physical objects and properties. Further, the non-deducibility of facts about the

former ($facts_{2.2}$) from facts about the latter ($facts_{2.1}$) is just what we would expect given what we know to be the nature of $facts_{2.2}$.

Before proceeding to my argument in Chapter 5 that physicalism does not require the deducibility of all facts from physical facts, let me briefly mention a way to resist the metaphysical version of the explanatory gap problem which is similar to the one I will present. This is a strategy developed by Brian Loar (1990) and Papineau (2002). Loar argues that knowing facts about conscious experience involves the possession of 'phenomenal concepts' (p. 84). These concepts (for example, the concept of what it is like to see red) can refer to the same properties as do physical or functional concepts (for example, the concept of being in brain state B). It is just that we do not know a priori that the two different concepts introduce the same properties. Thus, the non-deducibility of facts involving phenomenal concepts from facts involving only physical concepts does not entail that the phenomenal concepts introduce non-physical properties whose existence would entail the falsity of physicalism.

Nothing in this approach is incompatible with the arguments against the metaphysical version of the explanatory gap problem that I will present in the next chapter. Loar and Papineau argue (to translate their position into my terminology) that our inability to know $facts_2$ involving phenomenal concepts a priori from $facts_2$ involving physical concepts does not entail that the $facts_1$ underlying these $facts_2$ are different. This is the claim that I will be defending in Chapter 5.

Chapter Five: Does Physicalism Require Deducibility of all Facts from Physical Facts?

If physicalism requires that all facts be entailed a priori from the physical facts, then the existence of the explanatory gap means that physicalism is false. Using the distinctions among the different types of 'facts' put forth in Chapter 4, we can interpret the existence of an explanatory gap (the non-deducibility of facts about conscious experience from the physical facts) as the unexplainability (or lack of a priori entailment, or non-deducibility[22]) of $facts_{2.2}$ by $facts_{2.1}$. This chapter will complete my argument that this non-deducibility of $facts_{2.2}$ from $facts_{2.1}$ is not a metaphysical problem for physicalism. I will argue, against Jackson and Chalmers, that physicalism does not require the a priori entailment of all facts from the physical facts.

As I have already argued, the definition of physicalism set out above does not include, prima facie, any epistemological demands at all. All it says is that any possible world which is a minimal physical duplicate of the actual world must be a duplicate simpliciter of the actual world – in other words, there can exist nothing in the actual world additional to objects which are or are wholly constituted by the objects over which the theory that fundamentally describes all uncontroversially non-conscious objects quantifies. This version of physicalism makes no claim about what we must know, understand, or explain in order for physicalism to be true. It only claims that the objects and properties to which our descriptions of conscious experience apply are of the same type (physical) as the objects and properties to which our descriptions of non-conscious phenomena apply.

Given this, to show that the explanatory gap is a problem for physicalism, it must be shown that physicalism is committed to the a priori entailment of all $facts_2$, including $facts_{2.2}$, from the physical facts (which are $facts_{2.1}$). Both Jackson and

---

[22] Deducibility is Levine's condition for explanation (2001, p. 76). Jackson and Chalmers refer to the same basic notion as a priori entailment. I will use the two terms interchangeably in this chapter. Nothing turns on the difference, for this discussion. For my purposes, the notion of the deducibility of fact A from fact B, or the a priori entailment of fact A from fact B, is the conditional that if a subject knows fact B, then, in principle, the subject could come to know fact A without the character of her

Chalmers offer arguments that physicalism, as I have defined it, is committed to the a priori entailment of all facts from the physical facts (see, for example, Jackson 1998, pp. 81-83 and Chalmers 1996, p. 132). In this chapter, I will argue that these arguments do not succeed in providing a reason to overturn physicalism's prima facie lack of commitment to any epistemological claim.

The chapter has five sections. In section 5.1, I present the two-dimensional semantic framework that Jackson and Chalmers use to argue that physicalism is committed to the a priori entailment of all facts from the physical facts. In section 5.2, I present Jackson's argument, reinterpret it in terms of the different notions of 'fact' distinguished above, and argue that it is unsound. In section 5.3, I present and similarly reinterpret Chalmers's argument, and argue that it is unsound. In section 5.4, I argue that even if the objections I raise to the Jackson and Chalmers arguments in sections 5.2 and 5.3 do not succeed, the physicalist can respond to them by arguing that the two-dimensional semantic framework on which their arguments depend does not allow them to derive the epistemological commitment to which they want to yoke physicalism. In section 5.5, I present and evaluate an alternative way to argue against the Jackson and Chalmers arguments, proposed by Block and Stalnaker (1999).

5.1 The two-dimensional semantic framework introduced

Broadly speaking, the two-dimensional semantic framework is a method of semantic analysis in which for each word or phrase, there are two ways of determining what that word or phrase applies to, or what its extension is (in the case of a sentence, its extension is a truth value). There are many different versions of the two-dimensional semantic framework which, as Chalmers points out, differ widely both at and below their surfaces, causing them to yield different results in many areas of debate (2001). I will focus on the two-dimensional framework as it is used by Chalmers (1996) and by Jackson (1994, 1998), since it is their arguments involving the framework that I will be addressing. My explication here will be brief.[23]

---

experience providing any further justification for her belief in fact A beyond her belief in fact B (Boghossian and Peacocke 2000, p. 1).

[23] A detailed explication of the various forms and history of the two-dimensional framework is available in Chalmers's online paper, 'The Foundations of Two-Dimensional Semantics'

On the version of the two-dimensional framework that Chalmers and Jackson use, both ways of determining what a word applies to, or what its extension is, depend on a context, or state of a possible world. The difference between the two ways is that one depends on the context of the world in which the word is uttered, while the other depends on the context of the world in which the word is evaluated. The first way of determining what the extension is requires one to consider the world in which the word is uttered as the actual world. The second way of determining what the extension is requires one to consider the world in which the word is uttered as counterfactual.

So take the classic example of the word 'water' uttered at the possible world Twin Earth[24], in which XYZ fills the oceans and runs from the taps. If we determine the extension of 'water' according to the context in which it is uttered – Twin Earth – the extension is XYZ. To put it another way, the extension of 'water' is XYZ when Twin Earth is considered as actual. On the other hand, the context in which the utterance of 'water' at Twin Earth is evaluated is the actual world, where the word 'water' applies to H_2O. Thus if we determine the extension of 'water' according to the context in which it is evaluated, the extension is H_2O. To put it another way, the extension of 'water' is H_2O when Twin Earth is considered as counterfactual.

These two kinds of extension are determined by two kinds of intension – the primary intension and the secondary intension. The primary intension is a function that maps a word and a context considered as actual to an extension. The secondary intension is a function that maps a word and a context considered as counterfactual to an extension. On Jackson's and Chalmers's view, if a person understands or is a competent user of a word, he has access to the primary intension of that word, which means that he can determine what the word's extension would be at a context (were that context actual) if he is given all the facts about that context (Jackson 1998, pp. 82-83; Chalmers 1996, p. 58). So, again taking the classic example, if I understand the word 'water', then I know that at Twin Earth, where XYZ fills the oceans and runs

(www.u.arizona.edu/~chalmers/papers/foundations.html). See also the introduction to Szabo-Gendler and Hawthorne (2002, pages 39-55 especially).

[24] This example originates with Putnam 1975.

from the taps, the word 'water' would apply to XYZ. However, understanding the word 'water' does not entail knowing that its secondary extension (what it applies to at the actual world) is H_2O. One can understand the word 'water' without knowing that it applies to H_2O at the actual world. Unlike the primary extension of a word, the secondary extension depends on how the actual world is (since it considers other possible worlds as counterfactual), and a subject need not know all the facts about the actual world in order to understand words.

To enable possible worlds to be considered as actual, the two-dimensional semantic framework makes use of Quine's notion of 'centered worlds' (1969, p. 154). Centered worlds are ordered pairs consisting of a possible world and a particular viewpoint or perspective (e.g., a person) within that world. Centering is necessary for considering possible worlds as actual because if the considerer does not know the viewpoint of the user of the word in the context in which the word is used, he cannot determine what that word applies to in that context. To see this, imagine that the word is used at a world in which there is a planet like Earth and a very distant planet from Earth that is like Twin Earth. 'Water' could apply to either H_2O or XYZ at that world. One can only tell what the word 'water' would apply to if this context were actual if one knows what the perspective of the word's user is. (Chalmers 1996, p. 60). Thus, primary intensions map words and *centered* possible worlds to extensions.

5.2 Jackson's argument

Jackson and Chalmers argue, each somewhat differently, for the conclusion that physicalism is committed to the in principle a priori entailment of all facts from the physical facts. Though different in structure, their arguments both reach this conclusion by espousing a particular notion of what physicalism is (which basically accords with the one I have adopted) and a particular view of what it is to understand a word or concept[25].

[25] The normal use of the terms 'concept' and 'word' is that 'words' have meanings and those meanings are 'concepts', but Jackson and Chalmers run the two ideas together so that 'concepts', as they use the term, are like words in that they have meanings but are not meanings themselves (Chalmers 1996, p. 57; Jackson 1998, p 33). This mixing of terms should not matter for my discussion, and for ease of exposition I will adopt it in discussing their arguments.

First, let me present Jackson's argument, which can be sketched as follows:

A. If one understands a concept, one knows a priori what its extension is at a possible world if one knows the total context (all the facts) of that possible world.[26]

B. If physicalism is true, the total context of the actual world can be given in physical terms.

C. Thus, if physicalism is true, one can know a priori what the extension of any concept is at the actual world just by knowing the physical facts about the actual world.

D. Thus, physicalism is committed to the in principle a priori entailment of all the facts about the actual world from the physical facts about the actual world. (1998, p. 83).

Premise A in Jackson's argument states the two-dimensional semantic framework's view of what it is to understand a term or concept: understanding a concept means being able to determine what its extension would be in a possible world if given all the facts about that world. This is what it means to grasp the primary intension of the concept. Premise B is a claim about physicalism which is supposed to be derivable from the thesis that any minimal physical duplicate of the actual world is a duplicate simpliciter. C is supposed to follow from A and B, and the argument's ultimate conclusion, D, is supposed to follow from C. (I will have more to say about each of these steps.)

The step from A and B to C is valid: if it is true that understanding a concept means knowing a priori what its extension is at a possible world given the context of that world, and if the truth of physicalism implies that the context of the actual world can be given in physical terms, then indeed if physicalism is true, one should be able to know a priori what the extension of a concept is at the actual world just by knowing

[26] Jackson and Chalmers (2001, p. 323) note that this conditional is trivially true if the total context includes information about the concept itself. For example, if a subject were trying to determine the extension of 'water' in a given context, and if that context included the fact that the concept 'water' applies to the stuff in the oceans, the subject would know just from the context what 'water' applied to, but not because of any a priori grasp she had of the concept 'water'.

the context of the actual world in physical terms. The step from C to D, however, requires an additional, implicit premise:

> C_a: If one can know what the extension of any concept is at a world, then one can know all the facts about that world.

When this premise is made explicit, the argument is valid. But to determine whether the argument is sound we need to do three additional things. First, we must clarify what each premise means by 'facts' and evaluate the truth of the premises once this has been clarified. I will do this below by reinterpreting the argument in terms of the different notions of 'fact' discussed above. Second, we must determine whether and how premise B follows from the Jacksonian version of physicalism that I have adopted. I will address this question in the present section as well. Third, we must evaluate the two-dimensional semantic framework from which premise A is derived. I will take up this task in section 4.

Let me begin by reinterpreting each premise of Jackson's argument in terms of the different notions of 'fact' that I have defined. Premise A can be restated as follows:

> A'. If one understands a concept, and if one is given a set of $facts_{2.1}$ that tell one what all the $facts_1$ are at a given possible world, then one knows a priori what constituents of the $facts_1$ of that world are in the extension of that concept at that world.

I have interpreted Jackson's use of 'context' to mean a set of $facts_{2.1}$ which, together, tell what all the $facts_1$ are at a given possible world. This seems to accord with the rough definition of 'context' he gives: 'the relevant information about how things actually are' (1998, p. 83). 'Context' must be $facts_2$ because $facts_2$ are objects of knowledge, and Jackson talks about the context as something that a subject can know, or be given – as information. Thus he could not mean $facts_1$, or truth-makers, because truth-makers are not information that a subject can be given – rather they are collections of objects and properties which simply exist. Furthermore, Jackson must mean by 'context' $facts_{2.1}$, not $facts_{2.2}$, because the context is information that a subject can know about a counterfactual world. It is not clear how a subject could know $facts_{2.2}$ about a counterfactual world. As discussed above, $facts_{2.2}$ are facts that a

subject must have a particular experience of their subject matter in order to know. For example, to know that seeing red is like this, a subject must see red. If the subject matter is counterfactual – i.e., it does not actually exist – a subject could not have an experience of it, and without an experience of the subject matter, she could not know the fact. So if the subject's state of seeing red does not exist, the subject cannot experience being in that state, and so cannot know that being in that state is like this.

Next, we can restate premise B as follows:

B'. If physicalism is true, all the $facts_{2.1}$ needed to tell a subject what all the $facts_1$ are at the actual world can be given in physical terms.

B' follows from the Jacksonian version of physicalism that I have adopted. According to that version of physicalism, a minimal physical duplicate of the actual world is a duplicate simpliciter. This means that nothing exists or is instantiated in the actual world – there are no additional $facts_1$ – besides its physical objects and properties and what they constitute. Thus, all $facts_1$ of the actual world must be describable in physical terms, since by definition all physical objects and properties must be (at least in principle) describable in physical terms.

So I accept B' as an implication of this version of physicalism. For the moment, I will also accept A' (though in section 5 I will argue that understanding a concept in the context in which it is used requires more than just knowing its extension in that context given all the $facts_{2.1}$ about that context). Now the questions are whether C follows from A' and B', whether C_a is true, and whether D follows from C and C_a. To facilitate the answering of these questions, let me first restate the entire argument, adjusting the terms of C, C_a, and D to be consistent with those of A' and B':

A'. If one understands a concept, and if one is given a set of $facts_{2.1}$ that tell one what all the $facts_1$ are at a given possible world, then one knows a priori what constituents of the $facts_1$ of that world are in the extension of that concept at that world.

B'. If physicalism is true, all the $facts_{2.1}$ needed to tell a subject what all the $facts_1$ are at the actual world (or, the context of the actual world) can be given in physical terms.

C'. Thus, if physicalism is true, one can know a priori which constituents of $facts_1$ are in the extension of any of one's concepts just by knowing all the $facts_{2.1}$ needed to tell one what all the physical $facts_1$ are at the actual world.

C_a'. If one can know which constituents of $facts_1$ are in the extension of any of one's concepts, then one can know all the $facts_2$ about that world.

D'. Thus, physicalism is committed to the in principle a priori entailment of all the $facts_2$ about the actual world from the physical facts about the actual world, including $facts_{2.2}$.

One can see that C_a' is crucial for Jackson's conclusion that if physicalism is true then Mary must be able to deduce, from the physical facts she knows in the black and white room, that seeing red is like this. For without C_a', the argument only shows that if physicalism is true, then a subject, given sufficient $facts_{2.1}$ to tell him what all the physical $facts_1$ are at the actual world, should know a priori to which constituents of the $facts_1$ any concept he understands applies. But this does not itself entail the conclusion D'. C_a' is required for the step from A', B', and C' to D'.

But C_a' is false. Suppose that physicalism is true. If it is, then Mary (who knows all of physics in the black and white room) knows in the black and white room that the concept 'like this' (as used in the phrase 'seeing red is like this') applies to some physical property, call it brain state B. She knows what $fact_1$ her concept applies to. And yet she *still* does not know the $fact_{2.2}$ that seeing red is like this. This is because, as demonstrated by the knowledge argument, there are some facts that cannot be known without their knower having a particular experience, regardless of whether physicalism is true or not. Thus, just knowing which constituents of $facts_1$ are the extensions of all of one's concepts at a world does not entail knowing all the $facts_2$ about that world.

Of course, this argument that physicalism would not require Mary to know $facts_{2.2}$ a priori from physical $facts_{2.1}$ only raises the question of why it is that Mary cannot know the $fact_{2.2}$ that seeing red is like this in the black and white room. As I have said, my basic answer to this question is that $facts_{2.2}$ are not the sort of fact that can be a priori entailed by *any* other fact. Rather, they are facts that can be known

only if their knower has a particular experience: an experience directly of the subject matter of the fact. As I have also said, I will not offer an argument for the existence of $facts_{2.2}$ beyond its intuitive plausibility, which is demonstrated by the knowledge argument. My aim in this section has only been to argue that Jackson's argument depends on premise C_a' to show that physicalism requires the a priori entailment of $facts_{2.2}$ from the physical $facts_{2.1}$. And given the intuitive plausibility of the existence of $facts_{2.2}$, there is good reason to think that C_a' is false.

5.3 Chalmers's argument

Chalmers's argument that physicalism is committed to the a priori entailment of all facts (including $facts_{2.2}$) about the actual world from the physical facts about the actual world is different from Jackson's. Chalmers's argument can be sketched as follows:

1. The primary intension of a concept determines a property which is instantiated at every world at which the concept applies. (1996, p. 69).
2. The conceivability of a world entails the possibility of that world (p. 68).
3. If there is a minimal physical duplicate of the actual world that is not a duplicate simpliciter of the actual world, then physicalism is false.
4. Thus, by 1, 2, and 3, if we can conceive of a world which is a minimal physical duplicate of the actual world but lacks the property determined by the primary intension of a concept that applies at the actual world, then physicalism is false (p. 132).
5. One set of facts a priori entails another set of facts iff there are no two conceivable worlds that are identical with respect to the properties determined by the primary intensions of the one set of facts but different with respect to the properties determined by the primary intensions of the other set of facts (p. 70). [27]
6. Thus, if the antecedent of the conditional in 4 were false, there would be a priori entailment from the physical facts about the actual world to all facts about the actual world.

---

[27] 'Facts' here must mean $facts_2$, since $facts_1$ do not have intensions – rather, their constituents *are* extensions.

7. Physicalism requires that the antecedent of the conditional in 4 be false.
8. Therefore, physicalism is committed to a priori entailment from the physical facts about the actual world to all facts about the actual world.

As in Jackson's argument, Chalmers's conclusion follows from the definition of physicalism as the requirement that a minimal physical duplicate of the actual world be a duplicate simpliciter of the actual world and from the two-dimensional semantic framework. The argument is valid: 4 follows from 1, 2, and 3; 5 is definitional; 6 follows from 5; 7 follows from 4; and 8 follows from 6 and 7. To assess the argument's soundness, I will evaluate premise 1, since I accept premise 3, and since, as I have said, I will not address the truth or falsity of premise 2, the claim that the conceivability of a world entails the possibility of that world[28]. This premise can be used independently of the rest of the above argument to argue for the falsity of physicalism, as it is used in the zombie argument: this kind of argument is not my concern here.

Chalmers uses the premise that conceivability entails possibility to reach the conclusion that physicalism is false[29], not directly as in the zombie argument, but via premise 1. So if we accept for the sake of argument that conceivability does entail possibility, then the soundness of the argument depends on the truth of premise 1.

I will argue that premise 1 is false. My argument is informed by Lewis's argument, presented in Chapter 2, Section 2.1.3, that physicalism should be formulated as a thesis about natural properties. If we adopt Lewis's view that a physicalist thesis formulated in terms of duplication must be a thesis about natural, not abundant, properties – which seems quite plausible – then it is not clear that primary intensions, as defined within the two-dimensional semantic framework,

---

[28] See Chapter 3, section 3.2.

[29] To make the above into an argument that physicalism is false, all we need to do is add two steps:

9. We can conceive of a world that is identical to the actual world with respect to the properties determined by the primary intensions of our physical facts but different with respect to the property determined by the primary intension of the fact that, for example, seeing red is like this. This would be a 'zombie world', where the physical facts are the same as in the actual world, but there is no conscious experience – there is no property determined by the concept 'like this'.
10. By 6, 7, and 9, physicalism is false.

determine the relevant kind of properties – natural properties – such that if the instantiation of such properties were different between the actual world and some possible world, it would be shown that the possible world was not a minimal physical duplicate of the actual world. For the properties that Chalmers takes to be determined by primary intensions do not seem very much like natural properties.

Let me clarify what Chalmers means by 'properties determined by primary intensions'. He does not mean properties that are the *values* of primary intension functions given a world, but rather the properties that *are* the primary intension functions. He says:

> ...the primary intension determines a perfectly good property of objects in possible worlds. The property of being watery stuff is a perfectly reasonable property, even though it is not the same as the property of being H_2O. (1996, p. 132).

So the property determined by the primary intension of 'water' is the property of being 'the watery stuff', not the property of being H_2O, XYZ, or whatever the extension of 'water' is at any given world considered as actual.

But the property of being 'the watery stuff' is not a natural property. 'The watery stuff' is a placeholder description that Chalmers and Jackson use to represent whatever the function is by which subjects who understand the concept 'water' know a priori what its extension is at a possible world given the complete context of that world. Jackson and Chalmers acknowledge that there may be no explicit analysis of this function available. However, they maintain that subjects nonetheless know the primary intensions of their concepts, and hence must have access to such a function. (Jackson and Chalmers 2001, pp. 321-322). So being the watery stuff is the property of being the stuff identified by competent users of the concept 'water' as the extension of 'water' at a given world. This is not an intrinsic property, as it is a matter of what competent users of a word would say about objects whose intrinsic properties are already fixed before we make any judgment about them. Nor is it relevant to the causal powers of any object or property that instantiates it, since surely what we would say about a counterfactual world does not affect the causal relations that hold there. Perhaps one could argue that it does make for resemblances between objects: at the actual world, H_2O molecules resemble each other in that they all instantiate the

property of being stuff identified by competent users of the concept 'water' as the extension of 'water' at the actual world. But this does not make for any resemblance between objects beyond what the property of being H_2O makes for.

If Lewis's analysis is correct, then primary intensions would have to determine natural properties in order for Chalmers's argument to work. But it looks as though primary intensions do not determine natural properties. Thus, premise 1 of Chalmers's argument is false.

5.4 Against the two-dimensional semantic framework

I have argued that Jackson's and Chalmers's arguments that physicalism is committed to the a priori entailment of all facts – including $facts_{2.2}$ – from the physical facts do not succeed. However, as noted at the end of section 5.2 above, this still leaves the question of why it is that Mary, in the black and white room, could not come to know that seeing red is like this from knowledge of the physical facts and understanding of the concepts in 'seeing red is like this'. Why is it that, even if physicalism is true, the physical facts do not a priori entail the $facts_{2.2}$? I have suggested that the answer to this question is that $facts_{2.2}$ are a type of fact that cannot be entailed a priori from any other $fact_2$, simply because they can only be known through particular experience.

Jackson and Chalmers must reject this, because they use the two-dimensional framework to analyze our understanding of concepts. According to the two-dimensional framework, if Mary understands a concept and knows all the facts about the context in which it is used, then she knows what the extension of the concept is in that context. So the two-dimensional framework offers two possible explanations of why Mary does not know in the black and white room that seeing red is like this. One is that she cannot determine the extension of 'seeing red is like this' at a world given all the facts about that world. On this explanation, what Mary does not know in the black and white room, and what she learns upon leaving the black and white room and seeing red for the first time, is the primary intension of 'seeing red is like this'. The other possible explanation is that Mary does know the primary intension of 'seeing red is like this', but, having been given only physical facts about the world in which it

is used (the actual world), she does not know the total context of the actual world, and thus cannot determine the extension of 'seeing red is like this' at the actual world. On this explanation, what Mary does not know in the black and white room is the secondary intension of 'seeing red is like this'.

But the knowledge argument shows neither that Mary does not know the primary intension of 'seeing red is like this' in the black and white room, nor that she does not know the secondary intension of 'seeing red is like this' in the black and white room. First, I will explain why it does not show that Mary does not know the primary intension of 'seeing red is like this' in the black and white room. To simplify the argument, I will say that the fact which Mary does not know in the black and white room is that being in brain state B is like this. I make this simplification because I want to focus on whether Mary knows the primary intension of 'like this' in the black and white room. It may not be obvious that she knows the primary intension of 'seeing red' in the black and white room (as this may depend on whether she knows the primary intension of 'like this' in the black and white room, if part of what determines that something is the extension of 'seeing red' at a world is whether that something is 'like this'). But it is clear that she knows the primary intension of 'being in brain state B' in the black and white room. Let 'brain state B' be the physical state brains are in when the bearers of those brains have their retinae stimulated by light waves reflected from tomatoes, when they are disposed to say 'that's red', and so on. So the question is, does the knowledge argument show that Mary does not know the primary intension of 'like this' in the black and white room (or just that she does not know the primary intension of 'this')?

'This' is a demonstrative indexical concept. Presumably, Mary uses the concept 'this' competently in both speaking and thinking while in the black and white room. She knows what 'this' means; she knows how to apply it to various contexts. Roughly, she knows that the extension of 'this' in a given context is whatever is saliently indicated or ostended in that context. Thus, Mary knows the primary intension of 'this'. For knowing what the extension of a concept is given a context considered as actual is just the definition of knowing the primary intension of that concept.

The anti-physicalist might agree that Mary knows the primary intension of 'this' in the black and white room. If she does, the anti-physicalist might say, then her not knowing in the black and white room that being in brain state B is like this, despite knowing the physical context of the actual world, shows that the physical context of the actual world is not the total context, and thus physicalism is false, as Jackson's argument demonstrates. In this case, what Mary does not know in the black and white room – the $fact_{2.2}$ that being in brain state B is like this – is the secondary intension of 'being in brain state B is like this'.

But the knowledge argument does not show this either. The secondary intension of 'this' is the extension of 'this' at the actual world. If physicalism is true, the extension of 'this' as used in 'seeing red is like this' is, at the actual world, the property of being in brain state B. If physicalism is false, the extension might be some non-physical property. But whichever property is the extension of the concept 'this' as used in 'seeing red is like this' at the actual world, it is arguable that knowing which property is the extension does not alone suffice for understanding the concept as it is used.

The argument for this view relies on independent arguments of Gareth Evans (1985), which I will present it in the next paragraph. But before I do, note that if this view is correct, then even if Mary did know the secondary intension of 'this' in the black and white room, she would not know there that seeing red is like this. So if this view is correct, it cannot be that what Mary learns when she first sees a red tomato is the secondary intension of 'being in brain state B is like this'. And I have already argued that what she learns cannot be the primary intension of 'being in brain state B is like this', since she would have known that in the black and white room. Thus the two-dimensional framework would be shown to be inadequate to account for what Mary learns upon leaving the black and white room – the $fact_{2.2}$ that seeing red is like this.

Evans argues that knowing what referent[30] an indexical concept has in a context does not imply, as we might put it, knowing all facts expressed using the concept. Evans gives the example of the utterance 'today is fine', uttered on day *d*. Say that I know that the indexical concept 'today' refers to day *d* in the context in which the utterance was made. This alone is not sufficient for understanding 'today is fine' as uttered on *d*, Evans argues: what if *d* is the first day after the last lecture, and what if I think of *d* only as the first day after the last lecture? Then I understand the utterance of 'today is fine' as uttered on *d* as 'the day after the last lecture is fine'. But then I have not understood the utterance 'today is fine': I can only understand the utterance 'today is fine' if I think of *d* – the day that is the referent of 'today' in the context – as the current day. (p. 303).

A similar case can be made for the concept 'this', as used in the utterance 'seeing red is like this'. Does Mary, while in the black and white room, understand the utterance 'seeing red is like this', uttered by her (or her counterpart) in a possible world considered as actual where she is out of the black and white room and looking at a red tomato? It is clear that she does not: that is what the knowledge argument shows. And yet, as was just argued, Mary does know, or at least could in principle know while in the black and white room, both the primary and secondary extensions of 'this' as used in 'seeing red is like this'. As in the case of understanding the utterance 'today is fine', Mary must not only know the extension or referent of the utterance in a context to understand 'seeing red is like this', but she must also think of that referent in a certain way – as the experience that she is currently having.

Evans does not provide a definitive analysis of the notion of thinking of a referent in a certain way, but he suggests that the notion should involve the subject's bearing a certain relation to the referent:

> To give an account of how a thought concerns an object is to explain how the subject knows which object is in question. In the case of 'today', the subject, of course, knows which day is in question, but this knowledge at least partly consists in a disposition to judge the thoughts (which depend upon this knowledge) as true or false according to how things observably are upon that day which in no way rests upon his capacity to identify that day as meeting

[30] Evans uses 'referent' where I have been using 'extension'. Although they are different notions, it does not matter for my purposes here that I move between them.

> some antecedently given condition, but depends only upon his being alive on that day. Similarly, I should want to place in a central position in any account of what makes a man's thought concern a particular place in the way which is required for understanding sentences containing the term 'here', a knowledge of which place is in question which at least partly consists in a disposition to judge that thought as true or false according to how things observably are at that place – a disposition which he can have *vis-à-vis* just one place in the universe in virtue of his occupying it...(p. 304).

This suggestion parallels the suggestion I have made, that there are $facts_{2.2}$ which can only be known if the subject has a particular experience. In the above text, the relation that the subject must bear to the referent of the indexical concept is a spatiotemporal one. Evans also notes that the relation might be a causal one, such as that the subject is perceiving the referent (p. 314). These sorts of relations are different from the relation that would have to hold between a subject and the referent of a phenomenal concept for the subject to understand that concept. My suggestion is that in the case of Mary's utterance of 'brain state B is like this' at a world in which she is looking at a red tomato, even if she knows that in that context the referent of the phenomenal concept 'this' is such-and-such phenomenal properties, or such-and-such physical properties, she would not understand her utterance unless she was having the experience. This is analogous to Evans's argument that I would not understand the utterance in a particular context of 'today is fine' unless I occupied a particular spatiotemporal location in that context.

Note that the fact that the two-dimensional semantic framework specifies that primary intensions map *centered* worlds and concepts to extensions does not help Jackson and Chalmers here. The centering of a context allows the subject to determine what the extension of an utterance would be in that context if that context were actual. As discussed in section 5.1, if the context were not centered in this way, the subject would not be able to determine the truth value of the utterance (or the extensions of the concepts used in the utterance). However, this centering does not locate the subject *herself* in the context in which the words or concepts are used, such that she could bear the relation to the extensions of the concepts that are necessary for her to fully understand them as they are used in that context.

If the two-dimensional framework is not able to account for $facts_{2.2}$, as the application of Evans's arguments suggests it is not, then Jackson's and Chalmers's

arguments that physicalism requires the a priori entailment of all the facts about the actual world from the physical facts about the actual world cannot succeed. Jackson argues that if physicalism is true, then we can know what the extensions of all our concepts are just by knowing all the physical $facts_{2.1}$. This is because, according to the two-dimensional semantic framework, if we know the total context in which a concept is used and if we understand the concept, then we know what its extension is in that context. But if $facts_{2.2}$ are such that we cannot know a priori from knowledge of a total context what their extensions are in that context – if, as I have argued, such facts are not fully accounted for by the two-dimensional framework – then the non-a priori deducibility of what their extensions are from the *physical* context (which is demonstrated by the knowledge argument) does not show that physicalism is false.

Chalmers argues that physicalism requires a priori entailment of the $facts_{2.2}$ from the physical facts because without this a priori entailment the instantiations of the properties determined by the primary intensions of those $facts_{2.2}$ could be different between the actual world and a minimal physical duplicate of the actual world. But if the above argument is sound, then even if a property were determined by the primary intension of 'this' in the $fact_{2.2}$ that brain state B is like this, this property would be something like the property of being mentally indicated. And it would seem that given enough neurophysical information, we could indeed deduce that the property of being mentally indicated was instantiated at a world (at least, as much as we could deduce the facts about water at a world from the physical facts about H_2O at that world). Thus, the natural property determined by the primary intension of 'this' would not be one that we could conceive of a minimal physical duplicate of the actual world not having, any more than we could conceive of a minimal physical duplicate of the actual world not having water.

5.5 The Block and Stalnaker argument

I have argued that Jackson's and Chalmers's arguments that physicalism requires a priori entailment of all facts from the physical facts do not succeed because the two-dimensional framework on which their arguments depend does not account for the relevant kind of facts. Ned Block and Robert Stalnaker (1999) provide a

different way to reply to Jackson's and Chalmers's arguments, while accepting the two-dimensional framework.

The crux of their argument is the claim that standard, accepted identities between the extensions of everyday concepts (like 'water') and the extensions of physical/chemical concepts (like 'H_2O') do not have corresponding a priori entailments from the physical concepts to the everyday concepts. Thus, if the absence of a priori entailment from the physical facts to the facts about conscious experience demonstrates that conscious experience is not identical to or necessitated by the physical (as required by physicalism), then it can be demonstrated in the same way that water is not identical to or necessitated by H_2O, which is obviously false.

In arguing for the crucial claim that there is no a priori entailment in the standard cases, Block and Stalnaker present and attack the kind of argument that is used to support the claim that there *is* a priori entailment in the standard cases. They sketch an argument of Jackson's as their paradigm example of this kind of argument:

(a) It is a physical fact that the earth is covered 60% by H_2O.

(b) It is a physical fact that H_2O is the waterish stuff.

(c) It is a priori that water is the waterish stuff.

(d) Therefore, if one knows the physical facts and understands the concept 'water', one knows a priori that the earth is covered 60% by water. (Block and Stalnaker 1999, p. 12.)

I will present three types of argument that Block and Stalnaker make against Jackson's argument. The first is that there is no way of specifying the supposed a priori knowledge in (c) – that water is the waterish stuff – which reflects any a priori knowledge that we actually have by virtue of understanding the concept 'water'. No explicit a priori conceptual analysis is available for the concept 'water' to fully expand the shorthand 'the watery stuff'. (p. 16). The second argument is that (b) is false because it is *not a physical fact* that H_2O is the waterish stuff. They argue that the physical facts do not include the fact that there are no non-physical entities which are also 'waterish stuff'. Thus (b) is not a physical fact because it says that H_2O is the

unique waterish stuff, and yet the physical facts do not include this information. (p. 37). The third argument is that even if the property of being the waterish stuff is understood just as the primary intension of 'water', or the property of being the stuff that competent users of the concept 'water' would say is the extension of 'water' at a world, it still might not be a priori that water is the satisfier of the primary intension of water. Block and Stalnaker ask what we would say about a world in which there is both H_2O and XYZ, both playing the water role, or both being good candidates for a rational subject to identify as the primary intension of 'water'. Given such a case, they argue, we would not know a priori that water was the satisfier of the primary intension of 'water', because we would not know a priori what the primary intension of 'water' *was*. It would be indeterminate, Block and Stalnaker argue, what a rational subject should say about the extension of 'water' at that world. This means that it is not in general a priori that water is the satisfier of the primary intension of 'water', because we do not have an a priori method of determining what the satisfier of the primary intension of 'water' is at every possible world. (p. 23).

Jackson and Chalmers (2001) counter each of these arguments. To the first argument, they reply that even if there is no complete explication of 'the waterish stuff', there is still something we know a priori about 'water': we know what its extension is given a context. We know its primary intension. Thus, even if we cannot explicitly describe the function from centered possible worlds and concepts to extensions, we can still apply it, which means that (c), when understood to mean that water is the satisfier of the primary intension of 'water', is true: we can tell a priori that our concept 'water' applies to water in the actual world when given all the physical facts about the actual world. (pp. 321-322 and p. 337).

To the second argument, they reply that the thesis in question – that the facts about water are entailed a priori from the physical facts, given understanding of the concept 'water' – should be understood as the thesis that the water facts are entailed a priori, given understanding of the concept 'water', by the physical facts *plus* a 'that's all' condition, which says that the physical facts describe all that there is in that world. Block and Stalnaker object that this 'that's all' condition is neither part of nor implied by the physical facts: to say that it is would be to say that physics entails physicalism (p. 19). Jackson and Chalmers respond that the 'that's all' condition need not be

entailed by physics. They are just saying that the water facts are entailed a priori by the physical facts plus the 'that's all' condition, given understanding of the concept 'water'. (p. 342).

To the third argument, Jackson and Chalmers reply that it is not clear why Block and Stalnaker think that uniqueness is required for Jackson's argument to be sound (p. 340). If there were both H_2O and XYZ playing the water role at a world, then the argument might be something like this:

(e) It is a physical fact that the earth is covered 60% by H_2O and XYZ.

(f) It is a physical fact that H_2O and XYZ are the waterish stuffs.

(g) It is a priori that water is the waterish stuff.

(h) Therefore, if one knows the physical facts and understands the concept 'water', one knows a priori that the earth is covered 60% by water.

It seems that if we could know a priori from the physical facts what the different types of entities that constitute the extension of a concept were, we would just as easily know the facts about that concept a priori from the physical facts as we would if a single type of entity constituted the extension of that concept. Note that this argument only works if XYZ is a physical entity (i.e., one that can be described by physical facts). If XYZ is 'ghost water', it is not a physical fact that the earth is covered 60% by H_2O and XYZ. This kind of case would take us back to Block and Stalnaker's second argument.

But even if Jackson and Chalmers are right that Block and Stalnaker's arguments do not show conclusively that the facts about water are not entailed a priori from the physical facts, Block and Stalnaker's arguments do highlight the difficulty of showing that there is any contrast between the possibility of a priori entailment from the physical facts to the facts about water, and the possibility of a priori entailment from the physical facts to the facts about conscious experience. To see this, assume that Jackson's argument above ((a)-(d)) is sound (where being the watery stuff in premise (c) is understood to mean satisfying the primary intension of 'water'). Now take a similar argument about the concept 'pain', which Block and Stalnaker give:

(i) It is a physical fact that pyramidal cell activity was rampant in medieval prisons.

(j) It is a physical fact that pyramidal cell activity is the satisfier of the primary intension of 'pain'.

(k) It is a priori that pain is the satisfier of the primary intension of 'pain'.

(l) Therefore, if one knows the physical facts and understands the concept 'pain', one knows a priori that pain was rampant in medieval prisons. (p. 44).

This argument would be motivated: firstly, if it were discovered that pyramidal cell activity was rampant in medieval prisons – and if this were discovered, it would surely be among the physical facts, thus (i) would be true; and, secondly, if it were demonstrated that there was a strong correlation between experiences of pain and pyramidal cell activity, for in that case (j) would quite possibly be true. And surely, if it is a priori that water is the satisfier of the primary intension of 'water' (as in Jackson's (c)), then it must also be a priori that pain is the satisfier of the primary intension of 'pain' – so (k) is true. Thus, (l) follows, just as (d) does.

As Block and Stalnaker point out, it seems just as plausible that a rational subject given the facts about a world in which there is a strong correlation between pain and pyramidal cell activity would say that the extension of his concept 'pain' is pyramidal cell activity at that world, as that a rational subject given the facts about a world in which H_2O fills the oceans and runs from the taps would say that the extension of his concept 'water' is H_2O at that world. If there were an explicit conceptual analysis of 'pain', then we could say definitively whether or not pyramidal cell activity in that world fulfilled it, regardless of how strongly pyramidal cell activity was correlated with instances of individuals being injured, screaming, and so on. Or, if there were an explicit conceptual analysis of 'water' available, and not of 'pain', then the anti-physicalist could argue that it is the lack of explicit conceptual analyzes for phenomenal concepts like 'pain' that ensure that a rational subject could not identify pyramidal cell activity as the satisfier of the primary intension of 'pain'. But if we maintain that neither of these is the case, as Jackson and Chalmers do, then it is not clear what guidelines there are for the anti-physicalist to use in rejecting the notion that the primary intension of 'pain' determines that the concept's extension is pyramidal cell activity at such a world.

Further, take Chalmers's and Jackson's addition of the 'that's all' condition to the facts of physics. This condition also makes it difficult to see how the anti-physicalist can demonstrate a contrast in the possibilities of a priori entailment between the case of water and the case of conscious experience. They argue that if we know the full physical context, if we understand the concept 'water', and if we know that there is nothing in the world not accounted for in the full physical context of the world, then we know a priori (if we are rational) that H_2O is water.

But how could they argue that this does not also work for consciousness? If we know the full physical story, and we know that that's the complete story, then why would it not be rational to say that consciousness is whatever its physical correlate is? What other extension would better recommend itself as the satisfier of the primary intension of 'consciousness', in that case? Once again, if Jackson and Chalmers had argued that in the case of the concept 'water' there is some explicit analysis of consciousness which we can judge physical facts to fulfill, they could argue that the absence of such an analysis for the concept of 'consciousness' makes it impossible for us to judge any physical objects or properties to be the satisfiers of the primary intension of 'consciousness'. But they say that there need be no such analysis for 'water'. Thus they cannot argue that there need be such an analysis for 'consciousness' in order for us to judge physical objects and properties to be the satisfiers of its primary intension.

Chapter Six: Conclusion

I have argued that an explanatory gap exists, and that its existence is not a problem for physicalism. The explanatory gap, I have argued, amounts to the non-deducibility of facts about conscious experience from physical facts. Thus, for the explanatory gap to be a problem for physicalism would be for physicalism to be committed to this kind of deducibility.

In Chapter 5, I presented Jackson's and Chalmers's arguments that physicalism is so committed, and developed three options open to the physicalist to reply to these arguments. One way is by demonstrating that Jackson's and Chalmers's use of the notion of 'fact' is ambiguous because it does not take into account the different types of 'fact' that I distinguished in Chapter 4. When the notion is disambiguated the arguments do not succeed (as argued in sections 5.2 and 5.3). Another way is to argue that the two-dimensional framework cannot account for the type of facts that the knowledge argument shows are not deducible from physical facts. Therefore the framework cannot be used to argue that this non-deducibility shows that physicalism is false (as discussed in section 5.4). Since Jackson's and Chalmers's arguments depend on the two-dimensional framework, this suggests that their arguments cannot succeed. A third way is to accept the two-dimensional framework and follow Block and Stalnaker in arguing that if the non-deducibility of facts such as 'seeing red is like this' from the physical facts is a problem for physicalism about consciousness, then a similar non-deducibility of facts about water from physical facts is a problem for physicalism about water, which is clearly false (as discussed in section 5.5).

More would be required to fully develop these options, particularly the first two. These methods of dissolving the explanatory gap depend on the distinctions I have drawn between different types of facts, and I have left these distinctions largely intuitive. There are at least two questions unanswered. The first is whether there are indeed facts one can only know by having a particular experience – $facts_{2.2}$ – or whether such facts can in principle be learned by 'lessons': we just do not know how yet. I have argued that the knowledge argument demonstrates that there are $facts_{2.2}$, and that attempts to avoid this conclusion, such as the ability hypothesis, do not

succeed. This does not completely rule out the possibility that we might someday learn how to deduce facts about conscious experience from physical facts, but it certainly makes it seem implausible. Of course, if there really are no $facts_{2.2}$, and all facts can be learned by lessons, then the explanatory gap does not exist, and is not a problem for physicalism. My aim in this dissertation has been to argue that the explanatory gap exists, and that it is not a problem for physicalism.

The second question left largely unanswered is the question of how the distinction between $facts_{2.1}$ and $facts_{2.2}$ can be spelled out. What exactly does it mean for a fact to be knowable only via a particular kind of experience? I think that a distinction between $facts_{2.2}$ and $facts_{2.1}$ has great intuitive plausibility: the knowledge argument also makes this clear. Nonetheless, further development of this distinction would fill out the responses to Jackson and Chalmers that I have presented.

But while there is room in my arguments for further development, I hope that even in their current form they strongly suggest that the explanatory gap should be taken to be neither a great worry for physicalists nor a strong weapon for anti-physicalists. The explanatory gap describes the nature of facts about conscious experience: that they can only be known by the having of the experiences they are facts about. Physicalism is not a thesis about how we are able to know certain facts. The notion that certain facts can only be known through the having of certain experiences (on a physicalist picture, through the occurrence of certain neural events) is not in any way incompatible with the notion that all the objects and properties that make those facts true (the $facts_1$, as I have called them) are physical. Further, I have advanced a number of considerations against Jackson's and Chalmers's arguments that despite this prima facie compatibility of physicalism and the explanatory gap, physicalism nonetheless requires that there not be an explanatory gap. All of this indicates that physicalism will not be won or lost according to whether or not there is a way to bridge the explanatory gap.

Bibliography

Armstrong, D. M., 1978. *A Theory of Universals.* Cambridge: Cambridge University Press.

Beckermann, A., 2000. 'The Perennial Problem of the Reductive Explainability of Phenomenal Consciousness: C. D. Broad on the Explanatory Gap,' in T. Metzinger, ed., *Neural Correlates of Consciousness.* Cambridge, Massachusetts: MIT Press, pp. 41-55.

Block, N. and R. Stalnaker, 1999. 'Conceptual Analysis, Dualism, and the Explanatory Gap,' in *The Philosophical Review* 108, pp. 1-46.

Boghossian, P. and C. Peacocke, eds., 2000. *New Essays on the A Priori.* Oxford: Oxford University Press.

Broad, C. D., 1925. *Mind and its Place in Nature.* London: Routledge and Kegan Paul.

Campbell, K., 1970. *Body and Mind.* Notre Dame, Indiana: University of Notre Dame Press.

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

Chalmers, D., 2001. 'The Foundations of Two-Dimensional Semantics.' Online paper at www.u.arizona.edu/~chalmers/papers/foundations.html.

Chalmers, D. and F. Jackson, 2001. 'Conceptual Analysis and Reductive Explanation,' in *The Philosophical Review* 110, pp. 315-361.

Chalmers, D., 2003. 'Consciousness and its Place in Nature,' in S. Stich and T. Warfield, eds., *The Blackwell Guide to Philosophy of Mind.* Oxford: Blackwell.

Crane, T., 2001. *Elements of Mind.* Oxford: Oxford University Press.

Crane, T., 2003. 'Subjective Facts,' in H. Lillehammer and G. Rodriguez-Pereyra, eds., *Real Metaphysics.* London: Routledge.

Descartes, R., 1984. *Meditations on First Philosophy*, translated in J. Cottingham, R. Stoothoff, and D. Murdoch, *The Philosophical Writings of Descartes*, Vol. II.

Evans, G., 1985. *The Collected Papers of Gareth Evans*. Oxford: Clarendon Press.

Feigl, H., 1958. *The "Mental" and the "Physical"*. Minneapolis: University of Minnesota Press.

Fodor, J., 1974. 'Special Sciences,' in *Synthese*, Vol. 28, pp. 77-115; reprinted in J. Fodor, *Representations*, 1981, pp. 127-145.

Hempel, C. G., 1962. 'Explanation in Science and in History,' in R. G. Colodny, ed., *Frontiers of Philosophy*. London and Pittsburgh: Allen and Unwin and University of Pittsburgh Press, pp. 7-33. Reprinted in D-H. Ruben, *Explanation.* Oxford: Oxford University Press, 1993, pp. 17-41.

Horgan, T., 1993. 'From Supervenience to Superdupervenience: Meeting the Demands of a Material World,' in *Mind*, New Series, Vol. 102, Issue 408, pp. 555-586.

Jackson, F., 1982. 'Epiphenomenal Qualia,' in *Philosophical Quarterly* 32, pp. 127-136; reprinted in W. Lycan, *Mind and Cognition.* Oxford: Basil Blackwell, 1990, pp. 469-477.

Jackson, F., 1995. 'Essentialism, Mental Properties and Causation,' in *Proceedings of the Aristotelian Society*, New Series – Vol. 95, pp. 253-268.

Jackson, F., 1998a. *From Metaphysics to Ethics.* Oxford: Oxford University Press.

Jackson, F., 1998b. 'Postscript on Qualia' in his *Mind, Method and Conditionals: Selected Essays*. London: Routledge, pp. 76-79.

Kalderon, M., 2002. 'Old Facts Newly Known.' Unpublished paper.

Kirk, R., 1974. 'Zombies vs. Materialists,' in *Aristotelian Society Supplement* 48: 135-52.

Kripke, S., 1971. 'Identity and Necessity,' in *Identity and Individuation*, ed. M. Munitz. New York: New York University Press, pp. 135-164.

Levine, J., 1983. 'Materialism and Qualia: The explanatory gap,' in *Pacific Philosophical Quarterly* vol. 64, pp.354-361.

Levine, J., 1998. 'Conceivability and the Metaphysics of Mind,' in *Nous* 32, vol. 4, pp.449-480.

Levine, J., 2001. *Purple Haze.* Oxford: Oxford University Press.

Lewis, D., 1983a. 'Individuation by Acquaintance and by Stipulation,' in *The Philosophical Review* 92, pp. 3-32.

Lewis, D., 1983b. 'New Work for a Theory of Universals,' in the *Australasian Journal of Philosophy* 61/4, pp. 343-377; reprinted in D. H. Mellor and A. Oliver, eds., *Properties*. Oxford: Oxford University Press, 1997, pp. 188-227.

Lewis, D., 1988. 'What Experience Teaches,' in the *Proceedings of the Russellian Society,* University of Sydney; reprinted in W. Lycan, *Mind and Cognition,* Oxford: Basil Blackwell, 1990, pp. 499-519.

Loar, B., 1990. 'Phenomenal States,' in J. Tomberlin, ed., *Philosophical Perspectives*, Vol. 4. Atascadero: Ridgeview Publishing Company.

McGinn, C., 1991. *The Problem of Consciousness: Essays toward a Resolution.* Oxford: Basil Blackwell.

Melnyk, A., 2001. 'Physicalism Unfalsified: Chalmers's Inconclusive Conceivability Argument,' in *Physicalism and its Discontents.* Cambridge: Cambridge University Press.

Moore, A. W., 1997. *Points of View.* Oxford: Oxford University Press.

Nagel, T., 1974. 'What Is it Like to Be a Bat?' in *Philosophical Review,* 83:4, pp.435-450; reprinted in N. Block, O. Flanagan, and G. Güzeldere, eds., *The Nature of Consciousness*. Cambridge, Massachusetts: The MIT Press, 1997, pp. 519-527.

Nagel, T., 1991. 'Subjective and Objective,' in his *Mortal Questions.* Cambridge: Cambridge University Press.

Nagel, T., 1991. 'Panpsychism,' in his *Mortal Questions.* Cambridge: Cambridge University Press.

Nemirow, L., 1990. 'Physicalism and the Cognitive Role of Acquaintance,' in W. Lycan, *Mind and Cognition,* Oxford: Basil Blackwell, 1990, pp. 490-499.

Papineau, D., 1995. 'The Antipathetic Fallacy and the Boundaries of Consciousness,' in T. Metzinger, *Conscious Experience.* Paderborn, Germany: Ferdinand Schoningh, pp. 263-271.

Papineau, D., 2000. 'The Rise of Physicalism,' in M. Stone and J. Wolff, eds. *The Proper Ambition of Science.* London: Routledge.

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

Perry, J., 1977. 'Frege on Demonstratives,' in *The Philosophical Review* 86, no. 4, pp.474-497.

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

Putnam, H., 1975. 'The Meaning of Meaning,' in *Mind, Language and Reality*. Cambridge: Cambridge University Press, pp. 215-271.

Putnam, H., 1999. *The Threefold Cord: Mind, Body, and World.* New York: Columbia University Press.

Quine, W. V. O., 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Robinson, H., 1982. *Matter and Sense.* New York: Cambridge University Press.

Ruben, D-H., 1993. *Explanation.* Oxford: Oxford University Press.

Russell, B., 1927. *The Analysis of Matter.* London: George Allen and Unwin.

Snowdon, P., 2002. 'Knowing How and Knowing That: a Distinction and its Uses Reconsidered.' Unpublished Paper.

Stanley, J. and T. Williamson, 2001. 'Knowing How' in *The Journal of Philosophy* 98, pp. 411-444.

Szabo-Gendler, T. and J. Hawthorne, eds., 2002. *Conceivability and Possibility.* Oxford: Clarendon Press.

Van Gulick, R., 1993. 'Understanding the Phenomenal Mind: Are We All Just Armadillos?' in M. Davies and G. Humphreys, eds., *Consciousness*. Oxford: Blackwell, pp. 137-149.