Why Sensations Must be Neurological Properties: A Defense of the Identity Theory

by

©2013
Nicholas K. Simmons

Submitted to the graduate degree program in Philosophy
and the Graduate Faculty of the University of Kansas in partial fulfillment of the
requirements for the degree of Doctor of Philosophy.

Chairperson, John Bricke

John Symons

Thomas Tuozzo

Ben Eggleston

Michael Vitevitch

Date Defended: April 16, 2013

The Dissertation Committee for Nicholas K. Simmons
certifies that this is the approved version of the following dissertation:

Why Sensations Must be Neurological Properties: A Defense of the Identity Theory

Chairperson, John Bricke

Date Approved: April 16, 2013

ABSTRACT

In this dissertation, I defend the thesis that qualitative mental states known as qualia (e.g., tastes, feelings, pains) are identical to physical properties. In Chapter 1, I argue that qualia have a functional role in the world, and that is to facilitate non-automatic mental processes. In Chapter 2, I demonstrate how non-reductive accounts of the mind fail. In Chapter 3, I demonstrate how my reductive account fares better than similar accounts with respect to common and contemporary objections. In Chapter 4, I address arguments against any view like mine which seeks to understand qualia in a physicalistic framework.

CONTENTS

INTRODUCTION

We have a paradoxical relationship with the qualitative mental properties known as *qualia*. On the one hand, properties such as the *sensation of the color red*, the *taste of ice cream*, or the *painful prick of a pin* are intimately familiar. It seems obvious, for instance, how the taste of vanilla ice cream differs from the taste of chocolate ice cream. On the other hand, qualia are obstinately elusive, as we have yet to find a place for them in our understanding of the physical world. The natural place to look is the brain but, borrowing from Colin McGinn, we might despairingly wonder how "technicolor phenomenology" could possibly "arise from soggy gray matter" (1989, pg. 349). This problem has rightly come to be known as *the hard problem of consciousness* (Chalmers, 1996).

Despite the admitted difficulty in solving this problem, I aim to make some headway towards reconciling our understanding of qualia with our understanding of the physical world. In particular, I shall defend an account known as the *identity theory*, the idea that qualia are nothing "over and above" certain (yet to be determined) kinds of brain states. The theory, itself, is not new, as it dates back at least to the 1950s (Place, 1956; Smart, 1959). But, since the 1960s, it has been nearly unanimously rejected by the philosophical and cognitive science communities in favor of an understanding of the mind as (ontologically autonomous) software "realized" by the (non-essential) hardware of the brain (Putnam, 1967). As I shall argue, this was a mistake. Indeed, as it is becoming increasingly apparent that the dominant "mind as software" account has no place for qualia (Kim, 2005), looking back at the brain seems like the most reasonable course.

A BRIEF CHART OF THE TERRAIN

A first step towards determining how qualia fit into our understanding of the world is pinpointing what they do, or what they are for, in the larger scheme of things. So, the focus of Chapter One is sketching out an account of the causal function of qualia. Of course, some deny that qualia have any causal role in the first place. This might be because they are *eliminativists* and hold that a completed scientific understanding of the world will show us that there were never such things in the first place. Or it might be because they are *epiphenomenalists* and think that qualia exist but are superfluous byproducts of other physical processes. I argue that both of these views are untenable.

Having eschewed eliminativism and epiphenomenalism, I further argue that qualia play an essential role as an intermediary process between input from the world and output behaviors in biological creatures. Qualia are not necessary for certain kinds of behaviors. They are necessary, however, for a certain degree of *variability* in the kinds of behaviors we might see given any particular input. Why? The answer, as I argue, has to do with biology: our brains are such that they couldn't bring about this kind of variability in behavior without qualia. If this is correct, then we should seriously question whether or not we would be warranted in attributing consciousness to intelligent creatures made of non-biological hardware.

Determining the causal role of qualia only gets us so far in our attempt to understand them as physical phenomena. In Chapter Two, I discuss what we mean when we say that mental properties are physical properties in the first place. One might hold what has become the standard view of *non-reductive physicalism* and think that qualitative properties, though

they supervene on brain states, are nonetheless ontologically distinct. As I argue, this view is nothing more than a sophisticated kind of dualism; as such, it inherits the same kinds of problems we find with dualism, as it renders the attribution of qualia causally and explanatorily superfluous. The only other route we can take, then, is *reductive physicalism*, as Jaegwon Kim argues when making the case for functionalism (Kim, 1998). I argue, however, that functionalism, too, is a form of non-reductive physicalism and, so, it fails.

While non-reductive physicalism is conventionally attributed to Donald Davidson, I take an exegetical turn and argue that his actual view is a much weaker form of non-reductive physicalism that avoids the aforementioned problems, as talk of properties is avoided in favor of talk of predicates. I then extend this kind of thinking into an account of two kinds of explanations – *pragmatic* and *fundamental* – which accounts for the fact that a strong form of non-reductive physicalism is so attractive, despite its failure.

In Chapter Three, I argue for the type-identity theory. In particular, I argue that kinds (or types) of qualitative mental properties such as *pains* are to be identified with certain kinds of brain activity. This view goes back at least to U.T. Place and J.J.C. Smart in the mid-20th century, but, as I have mentioned, has since been rejected by the philosophical community. Since Smart's view is the closest to the one I defend, I focus on his account, discussing its virtues and vices. In the latter cases, I offer up my own responses to the standard objections. I then address more contemporary objections to the theory, arguing that we should reject the intuitive position of individuating qualia by their qualitative features, as described in a folk vocabulary. That is, I argue that we should reject the claim that anything

that *feels* like a pain, for instance, *is* a pain. Instead, qualia, like other natural kinds such as water, are individuated by their physical structure, as described in an explicitly physical vocabulary.

In Chapter Four – the final chapter – I focus my discussion on arguments against physicalist accounts of qualia, in general. I start by addressing David Chalmers' criticism of physicalism and the sophisticated semantic framework known as *two-dimensional semantics* on which he relies. I then turn to so-called *knowledge arguments* against physicalism, as espoused by Frank Jackson, which rely on some sort of epistemic difficulty to establish a metaphysical conclusion. Finally, I address the so-called *explanatory gap*, or the problem concerning how something like the brain can give rise to the rich, qualitative experiences we all enjoy.

CHAPTER ONE
WHAT ARE QUALIA FOR?

To say that qualia are physical properties is not to make the stronger claim that *everything* physical must have a causal role. If we commit ourselves to this, we have to discount, outright, the possibility of the existence of anything causally inefficacious such as broad mental content.[1] So, we should not follow Jaegwon Kim and accept what he calls 'Alexander's Dictum', the bi-conditional claim that "to be is to have causal powers" (1998) when we would do well enough to hold to the less contentious idea that *to have causal powers is to be physical*.

If qualia have causal powers, then we can be confident they are physical[2]. But how do we establish this? I shall argue in this chapter that they have causal powers because they are nomically necessary for a certain class of mental processes known as *controlled mental processes* (Shiffrin and Schneider 1984) in biological creatures. My discussion will start with critiques of those who hold that qualia have no causal role in the first place, such as the *eliminativists* and the *epiphenomenalists*. I shall end with a positive account: a sketch of what qualia do for us and creatures like us.

---

[1] For instance, the content of John's belief that a glass of water is front of him includes the glass of water, itself. But the glass of water doesn't motivate him to act; it is his belief about it that does.

[2] To be sure, there may be non-physical accounts of qualia, but methodologically speaking, we should opt for a physical account unless we are forced to think otherwise. In the third chapter, I shall elaborate on this methodological commitment.

§ 1: ELIMINATIVISM AND EPIPHENOMENALISM

'Eliminativism' does not refer to any one doctrine. Rather, we are eliminativists with respect to x, just in case we deny the existence of x. In terms of the usage of the word, we are generally considered eliminativists about x if there was first a widespread belief in the existence of x. For instance, many used to think that all organisms were infused with some kind of vitalistic life-force – or *élan vital* – which was responsible for their evolution and development (Dennett 1988). As we came to discover more naturalistic explanations for such phenomena, talk of such élan vital became eliminated from scientific discourse. So, nowadays, we are all eliminativists with respect to élan vital. In philosophy of mind, you find that 'eliminativism' usually refers to the denial of propositional attitudes and/or the denial of qualia. For obvious reasons, I shall focus only on arguments for the latter view, so any use of 'eliminativism' in this work will refer to the doctrine with respect to qualia, only. Since Daniel Dennett is arguably the most noteworthy figure arguing for this doctrine, I shall focus on his arguments.

§ 1.1: *DENNETT'S ELIMINATIVISM*

For Dennett, qualia do not exist because our concept QUALIA is incoherent. In this section I shall sketch out Dennett's argument for the nonexistence of qualia because of our conceptual incoherence; I shall then show that even if our concept is incoherent, it doesn't follow that there are no qualia.

In "Quining Qualia", Dennett first presents us with a series of thought experiments with the purpose of elucidating our concept of qualia (1988). Throughout this discussion, he identifies four essential properties of qualia, while trying to do justice to our folk intuitions:

1) they are *ineffable* in the sense that it seems impossible to explain, say, what the color red looks like to someone who is blind; 2) they are *intrinsic* features of our experience; 3) they are *private*, or not publicly observable; and 4) *they are directly or immediately apprehensible to consciousness* such that our knowledge of them is infallible (1988, pg. 523). So, according to Dennett, when philosophers use the term 'qualia', they are describing something containing attributes 1-4, essentially. Thus, in order for our concept of qualia to correspond with any feature of the world *x*, x must have these four essential properties.

After setting up the criteria that x must satisfy to be an instance of qualia, Dennett proceeds to work out a few cases where we intuitively think qualia are present, but, as he argues, they can't be, since at least one of the essential properties is missing. Consider, for instance, his fictional case of Chase and Sanborn, who work at Maxwell House as coffee tasters (1988, pg. 532). Chase tells Sanborn that, after six years of working there, he no longer likes the taste of Maxwell House. That is, the qualitative aspects of the taste are the same as they were six years ago, but he simply doesn't like it anymore. Sanborn, on the other hand, also reports that he has come to dislike Maxwell House, but because it no longer tastes the same to him.

The epistemic issue here, for Dennett, can be illustrated if we imagine further that Sanborn insists that he knows with certainty that the taste of the coffee at time *T1* – the beginning of his six years – differs qualitatively from the taste of the coffee at time *T2* – the end of the six years. What kind of evidence would lead us to accept or reject his claim?

We might think that we could settle the issue with neurological data. For instance, imagine we scan Sanborn's brain with fMRI at T1 while he is drinking Maxwell House coffee and discern that a region of his brain in state *B1* is strongly correlated with this qualitative mental experience *M1*. At T2 we discern that he is in brain state *B2* while drinking, where B2 ≠ B1. B2 is strongly correlated with *M2*. Perhaps, this data could tell us if it is true that M1 ≠ M2, but this is irrelevant. The issue at hand is the epistemic status of our introspection, not our neurological imaging techniques. So, inasmuch as our introspective reports are questionable, we must give up the fourth essential property, as our claims to knowledge of qualia are fallible. Since essential properties are necessary conditions for existence, it follows that qualia don't exist.

In response to being forced to give up this property, we might think we can bite the bullet and hold that infallible knowledge was never an essential property of qualia in the first place. In reply to this, Dennett says "The idea that one should consult an outside expert, and perform elaborate behavioral tests on oneself in order to confirm what qualia one had, surely takes us too far away from our original idea of qualia as properties with which we have a particularly intimate acquaintance" (1988, pg. 533). For the most part, this seems right. Dennett goes astray, however, with the 'too'. We might wonder why we can't just hold that we used to think qualia had one set of properties, but now we know they have another other set of properties?

The implicit argument against our response, as outlined above, relies crucially on a descriptive theory of reference, where it is held that we can only refer to a given feature in the

world if we have an accurate conception of that feature. Conversely, for such a theory, if our concept fails to capture or at least approximate the nature of that feature, then our concept fails to refer. So, if our concept SOUL carries with it a description of a set of properties including *a supernatural entity that interacts with the natural world*, but something can interact with the natural world if and only if it is natural, then SOUL has no referent.

One problem with this argument from descriptivism is that it derives an ontological conclusion from premises whose content is concerned primarily with the nature of language. In other words, Dennett argues that qualia don't exist because our talk about qualia is confused. Apart from mind-dependent properties and entities, the furniture of the external world stands independently of our conception of the world, so it is strange that something might not exist in virtue of some linguistic fact[3]. At best, it seems that the strongest conclusion we can get from Dennett's assumptions is a form of quietism, where we must hold that we just can't talk about qualia.

Even granting that Dennett is a quietist about qualia, this doctrine seems too strong. Consider, for example, that, in the past, we all had an inaccurate, prescientific conception of water. Let us assume, specifically, that some of us understood natural phenomena primarily in spiritual terms. With respect to water, we might say that the intension (subjectively construed) of the term 'water' included features such as *being the life-force of the world spirit*. Does it follow that, since there are no such things that have this feature, our utterances of 'water' failed to refer to anything in the world? The natural response to this, I think, would

[3] I am, of course, assuming that some kind of phenomenalism is false. I think this is a safe assumption.

be to point at an instance of water and say that we used to think of this kind of stuff one way, but we were incorrect. Instead of adhering to a descriptive theory of reference, we would naturally follow Saul Kripke and say that we were referring to water the whole time in virtue of our causal-historical relationship with it (1980).

We can concede to Dennett that our initial intuitions about qualia are wrong without concluding that this means there were never any to begin with. Just as we have refined our conception of water with an understanding of chemistry and physics, we should hope to refine our conception of qualia with a fully worked-out science of the mind, granting to Dennett that this conception cannot include the four aforementioned essential properties. Indeed, as we shall see later in this chapter, it looks like qualia are, in fact, *effable*. Further, in the third chapter of this dissertation, I shall make the case for the fallibility of our judgments about qualia.

§ 1.2: *EPIPHENOMENALISM*

Broadly speaking, the doctrine of epiphenomenalism is the idea that the world of the mental has no causal transactions with the physical world. More narrowly, we might distinguish two variants of epiphenomenalism: token and type (McLaughlin 1989). Token epiphenomenalism – otherwise known as *classical epiphenomenalism* – as a doctrine goes back at least as far as Thomas Huxley, who likened the relationship of mental events to physical events to that of steam whistles and steam engines (1874). With steam engines, all of the causal work that moves a train is at the level of the engine *E*, while the whistle *W* is simply a byproduct of the process with no function (at least it doesn't function to move the train). That is, the direction of causality in the case of the train is always E → W. (The analogy to

train whistles breaks down a bit, however, since whistles have causal powers over other things, whereas for the epiphenomenalist mental states have no causal powers whatsoever.)

Type epiphenomenalism, on the other hand, is the idea that mental events do have causal powers, but not in virtue of falling under mental types. That is, if a mental event *E1* causes a physical event *E2* it only does so in virtue of its physical properties, not its mental properties. So, for instance, the event E1 of John's being in pain may cause him to scream – event E2 – but it only does so in virtue of the pain's neurological properties, not its mental properties such as its qualitative character. In this section I shall sketch out a few common objections to epiphenomenalism, and then demonstrate how the epiphenomenalist might reply to these objections. I shall end on a methodological note, arguing that we should avoid the doctrine, despite no definitive arguments for its falsehood.

The first argument against epiphenomenalism is known as the *argument from introspection*. Introspectively, it certainly *seems* that we know that, for instance, it is the feeling of pain that causes us to scream out loud. How could one deny this? The problem with this argument, however, is that while we might observe a regular pattern of mental events, it doesn't follow that any one of these mental events causes the other. To hold that this is the case is to commit the fallacy of *post hoc, ergo propter hoc*. Further, the type epiphenomenalist may concede that mental events are related by cause and effect, but introspection fails at determining in virtue of what properties this occurs; it is the neurological properties, not the mental properties (Horowitz 1999, pgs. 425-426). So, we

must concede that the argument from introspection has no real force other than pointing out how extremely counterintuitive epiphenomenalism is.

Another objection to epiphenomenalism is that it runs into the problem of other minds (Jackson 1982). We seem to be warranted in holding that creatures like us have similar mental lives by appealing to the analogy that since certain kinds of behaviors are caused by certain kinds of mental states in us, it is reasonable that those same kinds behaviors in others are also caused by the same kinds of mental states. But, if epiphenomenalism is true, and mental states are causally inefficacious, we can't appeal to this analogy. Thus, it appears that we are not warranted in holding that others are conscious like we are. The epiphenomenalist has an obvious reply to this objection. Even though the mental causes nothing, kinds of mental states are certainly correlated with kinds of behaviors – and this correlation is guaranteed nomologically. So, all the epiphenomenalist needs to do in order to reply to this argument is give the following variant of the analogy: certain kinds of behaviors in us are correlated with certain kinds of mental states, so it follows that certain kinds of behaviors are also correlated with certain kinds of mental states in the case of others.

Finally, a relatively new objection to epiphenomenalism is known as the *argument from evolution* (Popper and Eccles 1977). The idea is this: it seems extraordinarily unlikely that creatures like us would have evolved to be conscious if consciousness weren't an adaptive trait. If consciousness is adaptive, then it serves some kind of function, and this necessitates its being causally efficacious. This objection might be devastating to the token epiphenomenalist, but the type epiphenomenalist may respond by holding that mental states

may be evolutionarily adaptive in virtue of their non-mental properties; this remains a possibility since not all traits are adaptive – consciousness might simply be a byproduct of other adaptive brain states (Horowitz 1999, pgs. 432-433).

So, there are no definitive arguments for the falsity of epiphenomenalism. Yet, epiphenomenalism is, by almost everyone's admission, extremely counterintuitive. If the doctrine is right, it is never the taste of ice cream that makes us say 'yum'; it is never the pain in the painful prick of a pin that makes us retract our hands. Rather, the causal work in our lives is done completely by the unconscious brain states on which our qualitative mental states supervene.

Though we have no philosophically compelling reason to think epiphenomenalism is *false*, I submit that as a matter of methodology we should avoid it until we have exhausted all other possible theoretical accounts of qualia. Firstly, at least when it comes to the mind, our intuitions seem to carry some sort of evidential status.[4] In the absence of any compelling reasons otherwise, then, we should think that our mental states are, indeed, causally efficacious (at this juncture, there isn't any evidence suggesting that epiphenomenalism is true). Indeed, there is philosophical precedent for the strategy of erring on the side of intuition in the absence of reasons to the contrary (Pryor 2000).

Secondly, if epiphenomenalism is true, we are left with an unsatisfactory lack of explanation as to why there are qualitative mental properties in the first place: it would simply be a brute fact that when you have physical states of the sort *X*, you have mental states

[4] For instance, our intuitive folk psychology posits the existence of things like memory – whose existence is vindicated later, empirically.

of the sort *Y*. That is, it would be a brute physical law that *X iff Y;* nothing further would explain this purported fact. While this might not seem problematic to those who hold the (Jerry) Fodorian view on special sciences – the idea that there are all sorts of basic "higher-level" laws not grounded in anything "lower" such as physics (1974) – if reductionism is shown to be the preferred view (something I shall argue later), it would be extraordinarily unlikely that there would be such a law in addition to the few basic laws of physics which, in principle, can explain everything else about the physical world.[5]

In conclusion, it is difficult to argue that epiphenomenalism is wrong, per se. This does not mean we have no reason for avoiding it, however. We can avoid the doctrine by laying out the theoretical advantages our account has over it. We might make an analogy with epistemology and liken the epiphenomenalist to the skeptic. We can probably never satisfy the skeptic on her own terms, but we don't need to; we just need to give the undecided good reasons for preferring the alternative over it. Likewise, it would be too demanding to require a proof of the falsity of epiphenomenalism; we just need to give good reasons for thinking qualia have causal powers.

§ 2: THE FUNCTION OF QUALIA

What are qualia for? The answer to this question might seem obvious to us. For example, after touching a hot stove, we retract our hands quickly and let out a scream. The touching of the stove is accompanied by the feeling of pain, so we think it is those qualitative properties that cause us to retract our hands. Yet, as the epiphenomenalist might point out,

[5] I will discuss this in Chapter Two.

the hand retracts before the feeling of pain sets in, because the nerve fibers carrying the pain signal transmit more slowly than the signal telling us to retract (Ramachandran and Hirstein 1998, pg. 439). So, if we are to determine the function of qualia, we must consult more than just our intuitions; we must look at the relevant empirical data from cognitive science. In this section, I shall sketch out an account of the function of qualia, trying to answer the question of what qualia are nomically necessary for. I shall then respond to a few objections, including one raised by Owen Flanagan and Thomas Polger, who argue that qualia are not nomically necessary for any kind of intelligent activity (1996). I shall end with a discussion of the implications of my account.

§ 2.1: *The Function of Qualia*

Let us return to the hot stove case from above. Recall that the qualitative experience of pain comes after the retraction of the hand. The epiphenomenalist might argue that this fact demonstrates that pain is simply a byproduct of the processes involved in this instance of stimulus and response. However, if we grant that pain is indeed an effect of some factor in this process, it does not follow that it is simply a byproduct with no causal powers. As I shall argue, it functions crucially to allow the agent choice, roughly speaking. For example, as V.S. Ramachandran and William Hirsten note, while your retraction response comes automatically, "…what you do about it [the pain] is flexible. You can put some medication on it, or you can run away from whatever caused it" (1998, pg. 424). More specifically, my working hypothesis is that qualia, in general, are properties of *controlled mental processes,* or

non-automatic mental processes[6] which function to facilitate an agent's ability to make choices. Behaviorally speaking, I shall call the ability to have multiple kinds of outputs given one kind of input or stimulus, *NARS* (non-automatic responses to stimuli). Since NARS are a behavioral result of controlled mental processes, we can infer that if NARS are present, then controlled mental processes are present.

The general theoretical framework under which I am operating is known as *dual process theory*. This idea – that there are two basic kinds of mental processes – goes back at least as far as Fodor's The Modularity of Mind, where he distinguished between what he called *domain-specific* and *domain-general* mental processes (1986). In the current literature on the subject, these processes are known as *system 1* and *system 2*, respectively (Evans and Frankish 2008). Processes in system 1 are carried out quickly, relatively effortlessly, unconsciously, and automatically. For example, when John (in English) asks Sally, a native English speaker, how she is doing, Sally doesn't have to try to translate the sounds into something meaningful; it comes to her automatically, even though her brain is carrying out a complex task to give her the result. System 2 processes, on the other hand are slow, effortful, conscious, and constrained by the resources of working memory. For example, if John (in French) asks Sally how she is doing, and Sally is struggling to learn French, she might take a while to understand what he is saying; she will have to recall and consciously apply the rules of the language she is learning. It is my contention that the qualia involved in system 2 processes are nomically necessary for them to occur. If this is right, we should expect that in

[6] I am restricting my quantification here to cover only processes carried out by biological creatures. The reason for this qualification will be evident later.

cases where qualia are absent, system 2 processes are lacking. Or, to put it another way, a biological agent who is given a particular stimulus will be able to make a choice concerning that stimulus if and only if he or she has a qualitative experience representing that stimulus. *No qualia, no choice.*

Luckily for us, there are actual cases where agents receive input from stimuli, yet lack the qualia that normally correspond to these stimuli. Restricting ourselves to visual qualia for the moment, let us examine specifically cases of what is known as *blindsight*, where persons receive visual input, but lack visual sensation (which, as it turns out, is not so lucky for them). As a result of an injury in the primary visual cortex – or V1 – those with blindsight believe themselves to be blind because they have no visual sensations. Yet, as the researchers who first stumbled upon this phenomenon note, they are nevertheless able to react to visual stimuli, though only to a certain degree (Humphrey 1974, Weiskrantz 1986). For instance, if an object is thrown at a blindsighter's face, she will move out of the way as an automatic response. Or, if a blindsighter is asked what kind of object is presented to her visual field, she will be able to guess correctly with a probability above chance, though she thinks she sees no object in the first place (Humphrey 2006).

Neurologically speaking, the reason these automatic functions are present, despite the lack of qualia, is because there exists a pathway separate from V1 running from the optic nerve to the brain stem that remains intact. In certain non-mammals, this pathway is all there is, as the development of V1 is a more recent evolutionary phenomenon. So, it seems

to follow that creatures like reptiles, despite their ability to react to visual stimuli, are blindsighters.

What all of this data on blindsight suggests is that vision is possible without qualia. More specifically, we might say that visual perception is possible without visual sensation. Modally speaking, the fact that the former is possible without the latter implies, as the psychologist Nicholas Humphrey argues, that they are numerically distinct kinds of things altogether, though we often conflate the two because of temporal coincidence (1992).

If seeing is possible without qualia, then what are qualia for? The answer lies in the fact that blindsighters are unable to perform anything other than automatic tasks, whether they are conditioned or hard-wired. There is little variability in the kinds of behaviors and mental processes that come as a result of visual stimuli. Indeed, in lieu of NARS, reptiles exhibit only automatic responses to stimuli in the world, such as snapping at small moving objects like flies. So, it is reasonable to infer that such lack of variability or ability to make choices is the result of the lack of qualia.

§ 2.2: *Possible Objections*

In terms of our account of visual qualia and V1, the epiphenomenalist may respond by arguing that we are not warranted in claiming that it is qualia that are necessary for controlled mental processes and NARS. Rather, they may claim that all we are warranted in holding is that it is V1 that is necessary; qualia simply come as part of the package. While this is certainly possible (at least epistemically possible), our account has certain explanatory advantages. First, it explains why we find qualia correlated with some kinds brain states and not others. The epiphenomenalist can give us no explanation as to why the brain

states involved in unconscious vision lack qualia, while other brains states involved in vision do have them; it just has to be a brute fact that that is the way it is. Second, our account tells us why there is consciousness in the world in the first place: it serves the important function of facilitating choice. Far back in our evolutionary history, creatures had no ability to make choices; now many do. The ability to make choices is arguably adaptive. Since qualia are necessary for choices, they are adaptive.

So far, the thesis I have been defending is that qualia are necessary for controlled mental processes and NARS. Flanagan and Polger, however, hold that approaches like mine that seek to identify what qualia are nomically necessary for are off the mark (1995). This objection stems from their commitment to the thesis of *conscious inessentialism* (CI): "...the view that for any mental activity M performed in any cognitive domain D, even if we do M with conscious accompaniments, M can in principle be done without these conscious accompaniments" (Flanagan 1984, pg. 309). For example, though one might effortfully and consciously play a particular segment of a musical piece on guitar, it is nomically possible to play it unconsciously; this is evidenced by the fact that this generally happens for most musicians after practicing for a significant amount of time.

If CI is right, then the existence of qualia is an accident. To put it another way, qualia are accidental properties of some particular set of non-qualitative mental properties. In light of this, Flanagan and Polger contend that if our evolutionary history had panned out

a different way, there might have existed beings functionally equivalent to us, but who lack qualia (1995, pg. 3). That is, if CI is true, functional zombies[7] are nomically possible.[8]

As a matter of logic, however, the truth of CI does not imply that functional zombies are nomically possible. Even if we grant that any *particular* mental process may (nomically) be carried out without qualia, it does not follow that the conjunction of all particular mental processes we normally find being carried out by the average conscious person at any given time may be carried our without qualia. That is, even if it is possible that x, y, or z may be carried out unconsciously, it does not follow that x, y, and z may all be carried out unconsciously at the same time. For example, just because we might be able to drive particular stretches of an automobile trip without conscious awareness, it does not follow that we may drive an entire trip unconsciously.

The fact that the truth of CI does not imply that functional zombies are possible does not hurt Flangan and Polger's account in general, however. If CI is true, our entire methodology for understanding qualia is flawed. Now, I have already sketched out a plausible alternate account concerning how qualia might have some essential function, but

---

[7] I use 'functional zombies' to differentiate these zombies from the kind of zombies David Chalmers talks about, which we will discuss later in this dissertation. Briefly, the possibility of functional zombies does not imply that a *molecule for molecule* duplicate of a conscious entity may be unconscious. For now, we are only concerning ourselves with functional duplicates.

[8] To clarify some possible worries, the purported possibility of zombies in this sense does not imply the falsity of physicalism. Nor does the truth of CI entail epiphenomenalism. Just because it is possible for any kind of mental activity or behavior to be carried out without qualia, it does not follow that qualia don't carry them out, actually. For instance, consider that flying for birds is carried out by their wings. It is certainly nomically possible that something other than wings could have carried out this behavior, but this possibility does not entail that the wings, as a matter of fact, are causally inefficacious with respect to flying.

the truth of this account is dependent on a further question concerning how qualia evolved. If qualia have an essential function, then then they must be adaptive traits.

Flanagan and Polger, however, hold that we should not discount the possibility that qualia are non-adaptive and thus not essential for any given ability (1997). To illustrate the claim that qualia are inessential, Polger cites the fact that a bird's ability to fly, though facilitated by their wings as a matter of fact, might (construed nomically) have been carried out by something radically different (2007, pg. 15). Indeed, we might point to the fact that the ability to move, in general, may be carried out in many different ways: walking, slithering, jumping, just to name a few methods of locomotion. In light of this, Polger argues that consciousness would be a very special if it were necessary for some kind of ability (2007, pg. 3).

I contend, however, that holding that qualia are essential for x does not make them special, since there are other traits we may find in nature that are like this. For instance, the eye is a physical trait we see evolve time and time again, despite divergent evolutionary histories. We find it not only in humans and other mammals but in octopi and fish as well. The fact that we see the eye in such varied creatures suggests that eyes are nomically necessary to see, at least for biological creatures.[9] In sum, I contend that the instances of qualia we find in nature are like the instances of eyes we find in nature: the same solution to the same problem.

[9] You might think that bats can see without eyes, so eyes aren't necessary to see. But, even though we talk as if it is a kind of seeing, we would be equivocating to hold that it is the same kind of seeing we experience.

As I stated before, if qualia serve an essential function, then they must be adaptive. Though I have countered Polger's argument for qualia not being adaptive, if we can't give positive reasons to think that qualia are indeed adaptive, then there isn't much reason to favor my account over his. So, how shall we settle this? The best we can do at this juncture is offer up a hypothesis and judge its relative merits. What I have in mind is what I shall call the *material constraint hypothesis* (MCH) concerning the existence of qualia: *certain material constraints on our biology/neurology are such that a certain class of adaptive mental processes (and resulting behaviors) cannot be carried out without qualia.* MCH does a great deal of explanatory work, as we shall see.

First, if, say, dolphins are conscious, MCH would explain why they are, despite their having markedly different evolutionary histories from ours, as qualia are necessary for the having of controlled mental processes and NARS. If, on the other hand, CI is right, we would only be able to avail ourselves of a causal-historical explanation, where we must accept that it is simply coincidental that qualia just happened to arise not once, but twice in two respective evolutionary histories. While such an explanation might be sufficient in cases of genuine historical accident, it is unlikely that qualia would have evolved multiple times in different histories if they were not necessary for controlled mental processes. If CI is right, rather, we should expect dolphins to be unconscious, since kinds of non-qualitative brain states far outnumber qualitative ones.

Second, it seems reasonable that qualia are present in some animals, but not others. For instance, one might reasonably assume that bees aren't conscious, while dogs are. If this

is right, why would it be the case? MCH would give us a criterion for determining which creatures have qualia and which creatures don't. When we encounter beings with NARS, we can infer that they are conscious, and we can say why. Those defending CI, however, might reply that they, too, have a criterion because historical constraints on evolution will guarantee that the same kinds of brain states will appear in an evolutionary history of one species, and perhaps amongst several species, given common ancestors way back. The problem with this response is that it seems that the existence of qualia is a relatively recent evolutionary phenomenon, so historical constraints should have little bearing on the issue. For instance, it would be unlikely that the presence of qualia in both humans and dolphins can be explained by appealing to some common ancestor.

Finally, MCH would explain why the doctrine of functionalism fails when it comes to qualia. Briefly, functionalism is the idea that, to use Block's words, "mental states are constituted by their causal relations to one another and to sensory inputs and behavioral outputs" (1996, pg. 1). So, a kind of mental state such as a belief is not a kind of brain state as the (type) physicalist would have it, nor is it a kind of behavioral disposition as the behaviorist would have it. Rather, what makes x a mental state M, is just what x does, or the functional role it performs (just as whatever makes an object a chair is the function it performs). In this way, mental states are likened to software, since what is essential to software is the function it performs, not the hardware performing the function. Qualia, however, by almost everyone's admission stubbornly elude functional reduction.[10] Given

[10] This will be argued for in the next chapter.

MCH, we can say that functionalism fails in this respect because qualia, to use the computer metaphor, are part of the hardware side, but not part of all kinds of hardware. In our case, for our brains to carry out certain functions, qualia must be present because of certain material constraints. Such constraints needn't be present for the same functions to be performed by different hardware.

Given that we have admitted that the aforementioned material constraints needn't apply in cases of non-biological hardware, Flanagan and Polger might respond that I have not shown that CI is false after all, since, in principle, we might encounter robots with NARS but without qualia. We may reply in a couple of ways. First, if CI is quantifying over kinds of mental processes, we can concede that, yes, it *might* be the case that NARS may be carried out without qualia, but only by synthetic creatures. When it comes to biological creatures, it is not nomically possible for NARS to be carried out without qualia. So, the possibility of functionally equivalent zombies isn't so disconcerting, since it is not possible for humans to be zombies. Also, given that CI is false with respect to biological creatures, we are not barred from holding that qualia are adaptive; nor are we barred from determining their function. Second, if CI is taken to quantify over particular instances of mental processes, it is indeed false. That is, if we construe CI to mean something like *for any particular creature C: any of C's mental activities M performed in any cognitive domain D can in principle be done without conscious accompaniments, even if he or she does M with conscious accompaniments*, then it is false. So, construed the former way, CI lacks any of the novel implications it was intended to have. Construed in the latter way, CI is false

§ 2.3: *IMPLICATIONS*

If what I have argued for so far is correct, qualia exist, in some sense, because of a *flaw* in our hardware. Most mental processes are carried out by system 1 unconsciously, quickly and efficiently. System 2 processes, on the other hand, take a significant amount of time and use a non-negligible amount of resources or energy (Evans and Frankish 2008). If all mental processes could be carried out by system 1, they probably would be, since, all things being equal, natural selection would favor those creatures who could perform NARS efficiently over those who must expend significant amounts of energy to perform these tasks - slowly, at that. But this isn't the case, and it seems not to be a historical accident.

Whether something counts as a flaw or a feature is in the eye of the beholder, so we might want to consider qualia a feature of biological processes. If this is right, we might wonder if other beings such as robots might have the same "flaw". Must a robot be conscious in order to have controlled mental processes or NARS? It is not metaphysically necessary that this be the case, since functional equivalence between a person and a robot does not guarantee the equivalence of qualitative mental properties. So, we are not in a position to answer this question at this juncture. There might be a way to answer this question in principle, however. Consider the case of vision and V1. If visual qualia are located in V1, we could do the following to test whether or not a different physical medium may be conscious. First, we could shut down V1 from functioning in some willing person with a method known as transcranial magnetic stimulation (TMS), which interferes with neuronal activity. Then we could temporarily replace V1 with an artificial functional duplicate via a cable of

neurons and some sort of interface.[11] Finally, we would try to get a verbal report from the person to determine if they were experiencing visual qualia. An answer in the negative would either confirm our hypothesis or (at least) corroborate it (in the sense that it has not been falsified), depending on one's philosophical views concerning hypothesis testing.

One might worry that we wouldn't be able to determine much from such a report, since zombies would say the same thing. Given that all other sensory modalities (touch, smell, etc.) would be functioning, however, these epistemic worries should be assuaged. An additional epistemic worry one might have would be the possibility that V1 is merely a necessary condition for the experience of visual qualia. If this were right, it wouldn't follow that the experiences of such qualia would be located in V1. So, our functional V1 replacement test wouldn't tell us much because V1 would just be transmitting information to another area in the brain where the qualia are located. Studies using fMRI suggest, however, that visual qualia in general are located in V1 (Le Bihan, Turner et al. 1993).

Another implication of this view concerns our perception of time: namely, the fact that time seems to speed up as we get older. While there may be a number of factors involved in this phenomenon, my view suggests that this stems from the fact that as we get older, our everyday activities tend to become more automatic. I recall that when I was first learning to play the guitar, practicing for hours was a kind of endless drudgery. Now, since I have mastered the instrument, I find that I can practice for hours and be surprised that so

[11] For more discussion on how this might be set up see Ramachandran and Hirstein, 1998, pg. 432.

much time has passed by. If the account I have sketched out is correct, the reason for this is that guitar playing went from being a primarily system 2 process to being a system 1 process.

CONCLUSION

In this chapter I have argued for the intuitive view that qualia play a causal role in the world, in general, and an adaptive role for us, specifically.

In Section 1, I began by addressing arguments against this view, focusing on eliminativism and epiphenomenalism. For the eliminativist, talk of 'qualia' is nonsense, and so we should eliminate it from our discourse. As we have seen, arguments of this stripe rely on the idea that our philosophical intuitions with respect to qualia are incoherent, thus our use of 'qualia' fails to refer to anything in the world. It seems implausible, however, that a lack of conceptual coherence with respect to x implies that we can never refer to x.

The epiphenomenalist, on the other hand, agrees that there are qualia but holds that these properties have no causal role in the world. The problem with epiphenomenalism is that adherence to the doctrine commits us to holding that the relationship between the mental and the physical is primitive. That is, for the epiphenomenalist, mental states are correlated with physical states, and this is a basic law of the universe. As I have argued, this is not something we want to accept until all other options have been explored.

In Section 2, I gave a positive account of the role of qualia in the world. Contra Flanagan and Polger, qualia are, I argued, essential properties of controlled mental processes in biological creatures: mental processes – beyond our genetic and environmental programming – of the kind that we utilize in order to make choices. In response to the

question of *why* qualia are essential properties of controlled mental processes, I advanced *MCH* – the "material constraint hypothesis" – or the idea that our biological material is flawed in such a way that only one configuration yields controlled mental processes. This configuration happens to be the one from which consciousness "emerges". This might seem unlikely, but there are precedents in traits like the eye. Just as biological creatures require eyes for vision, we require qualia for controlled mental processes.

A couple of notable implications follow from this account of qualia. Firstly, just as non-biological creatures don't require eyes for vision, we have no reason to think that non-biological creatures exhibiting controlled mental processes must be conscious. This puts a damper on any hopes of transcending our biology entirely. Secondly, we have a tentative explanation for why time seems to go faster the older we get. Since more and more of our mental processing becomes automatic as we age, more and more of our mental processes become unconscious. If this phenomenon is disconcerting, we can slow it down by actively engaging in effortful mental tasks.

CHAPTER TWO
NON-REDUCTIVE PHYSICALISM AND QUALIA

Reductionism used to be a prevailing view in philosophy. In this vein, philosophers – particularly the logical positivists – working in the period between the 1920s and 1960s pursued the ideal of the unity of science, or the idea that all sciences are reducible, in principle, to physics. For the most part, nowadays, the unity of science is no longer considered ideal. Along these lines, the term 'reductionism' carries a mostly pejorative connotation – the idea being that pursuing reduction is too limiting or eliminativistic. We can trace this shift in attitude to the 1960s and 1970s, primarily with the influential works by Hilary Putnam (1967) and Jerry Fodor (1974). The alternative – now orthodox – view coming out of this tradition is known as non-reductive physicalism (henceforth known as *NRP*). For the proponent of NRP like Fodor, science is "disunified". That is, it is not the case that all sciences are reducible to physics, even in principle. Rather, the special sciences such as psychology are completely autonomous in the sense that understanding psychological phenomena comes completely independently of our understanding of the underlying physics.

Despite the broad consensus that NRP is the correct view, there is no consensus concerning how we should construe it. So, in this chapter I shall attempt to clarify how we should interpret this doctrine. In particular, I shall sketch out a distinction between two strains of NRP: strong and weak (henceforth known as *SNRP* and *WNRP*, respectively). At a first pass, those adhering to SNRP hold that mental properties are not reducible to physical

properties talked about in any explicitly physical science such as neuroscience, and the implication of this is that mental properties must be "higher-level" properties, such as the property of *being a lamp* (a property not identical to any explicitly physical property). As such, those adhering to SNRP also hold that the nature of the mental cannot, in principle, be explained by theoretical accounts at the "lower-level" such as at the level of neurology or physics. Those adhering to WNRP, on the other hand, can grant the irreducibility of mental properties in a *sense* (in the sense that we cannot identify mental properties with properties found in an explicitly physical science), but hold that, at the end of the day, there is, indeed, a "lower-level"[12] explanation of the mental to be found. The presence of such an explanation, however, for those adhering to WNRP, is not at odds with the presence of an additional "higher-level" explanation. Further, those adhering to WNRP hold that, even though particular mental kinds might not be identified with explicitly physical kinds, mental properties are nevertheless nothing "over and above" physical properties. We shall see what this means, exactly, in Section 2. It is my contention that WNRP is the only viable view. In this chapter I shall show why this is the case, starting with why SNRP in an unworkable account of the mental.

Before we get to the next section, let me make a brief stylistic note. Given that our overall project concerns qualia, I shall focus my discussion on NRP with respect to mental

[12] I put scare quotes here to designate the fact that, on my account, talk of different explanatory levels is not meant to imply that there are different levels of reality, ontologically speaking.

properties.[13] Further, it will be important for us to discuss this doctrine with respect to the mental, in general (including intentional states), before we get to qualia. The rationale for this move is the plausible idea that determining the role of qualia within the framework of the mental, in general, will help us see how qualia fit within a physicalistic framework.

§ 1: STRONG NON-REDUCTIVE PHYSICALISM

In this section, I shall chart out the origins of NRP and track its evolution into SNRP. I shall then examine SNRP in light of two closely related problems concerning the difficulty its adherents have in reconciling the fact that physics does either all of the causal work, or all of the explanatory work with their commitment to the idea that mental properties are distinct from physical properties. The first is the *causal exclusion problem,* the second is the *explanatory exclusion problem*.

§ 1.1: *DAVIDSON'S ARGUMENT FOR NON-REDUCTIVE PHYSICALISM*

Though we can trace non-reductive sentiments back decades farther, the origin of the standard formulation of NRP dates back to Donald Davidson's influential paper "Mental Events" (1970). In this paper, Davidson is concerned with reconciling a commitment to physicalism with the apparent fact that there are no strict laws governing the mental[14]. Up until this time, most philosophers thought that if the mental is physical, there must be psychophysical laws connecting the mental to the physical. For instance, it was thought that sensations could be physical only if there could be laws such as *an agent is in a state of pain if*

[13] As opposed to, say, economic properties or astronomical properties.

[14] It is important to note that Davidson is concerned with propositional attitudes, and not qualia. So, I am giving a broad treatment of his account.

and only if the agent's c-fibers are firing. Such laws would allow intertheoretic reduction from our folk theories concerning mental states to an explicitly physical theory such as neuroscience and, in turn, would allow us to see how the mental could be nothing over and above the physical. Davidson argued, on the contrary, that psychophysical laws are not necessary for us to hold that the mental is physical. The view that emerges from this consideration is what he calls *anomalous monism*.

Specifically, Davidson is concerned with reconciling the following three plausible but apparently inconsistent principles:

1. The principle of causal interaction: *at least some mental events interact causally with physical events.*
2. The principle of the nomological character of causality: *events related by cause and effect fall under strict laws.*
3. The anomalism of the mental: *there are no strict laws on the basis of which mental events can be predicted or explained.*

So, mental causation requires strict laws, but there are no strict laws governing the mental. How can this work? Davidson resolves this problem by noting that, at least in his view, laws are a linguistic phenomenon (1970, pg. 141). So, to say that there are no laws governing the mental is to say that there are no laws with mental predicates. For him, rather, the mental is still governed by strict laws, but only as described in a physical vocabulary. Mental events, then, are physical events, but the mental is distinct from the physical because our

explanations of these events in mental terms are not reducible to explanations in physical terms.

In the literature, it has become standard to formulate Davidson's view in terms of properties (Kim 2003).[15] From this, we get the view that a given particular mental event is an event *e* with mental properties and physical properties. For instance, if John is having a belief *B*, we have event e with physical properties such as *being a particular brain state* and mental properties such as *being an intentional state*. This form of non-reductionism is known as *token-physicalism*, which is the idea that mental events are physical events, but the mental is not physical in the sense that no particular mental kinds may be identified with particular physical kinds. To put it another way, *tokens* of mental events are identical to tokens of physical events, but *types* of mental events are not identical to types of physical events.[16]

Despite the seeming inability to reduce the mental to the physical, the relationship between the two, Davidson suggests, is that of supervenience (1970, pg. 141). Mental properties supervene on physical properties, in the sense that there can be no change in the former without a change in the latter. So, mental properties are strongly dependent (metaphysically, not causally) on physical properties, without being reducible to them.

§ 1.2: *The Causal Exclusion Problem*

Davidson's view, as we have construed it thus far, has a few problems. To sum up one problem in Jaegwon Kim's words, "Supervenience is not a theory" (1998, pg. 9). That is,

[15] As we shall see, the standard formulation is incorrect. But we shall proceed this way to get a lay of the land.

[16] This account with respect to identity is known as the *token identity theory*.

to say that the mental supervenes on the physical doesn't tell us much about the relationship between the two. Consider, for instance, that facts concerning water supervene on facts concerning H2O. It would be nice if this fact, itself, had an explanation; that is, we might wonder *why* water supervenes on H2O. In this case, the explanation is simply that water *is* H2O. In the absence of some further fact grounding this relationship, however, the relationship between the mental and physical seems to be primitive. So, the answer to the question concerning why the mental supervenes on the physical is simply "It just does."

Further, Davidson's view, despite his concern to establish the causal efficacy of the mental, seems inevitably to lead to type-epiphenomenalism (McLaughlin 1992). For Davidson, a mental event *e* causes a physical event *e** in virtue of being connected by a strict physical law. So, let's say that e has properties *M* and *P*, and e* has property *P**. The mental event e causes e* in virtue of the causal relationship between P and P*, not in virtue of any relationship between M and P*. Mental events *qua* mental, then, are not causally efficacious. They are only causally efficacious *qua* physical.

The conjunction of the supervenience problem with the problem concerning causal efficacy, along with a few other plausible commitments, gives us what is known as the *causal exclusion problem*, concerning how difficult it is to reconcile the commitment of distinctness of the mental from the physical with the commitment that the mental is causally efficacious. Indeed, as we shall see, this is a problem not just for Davidson's view (as we have thus far construed it), but for any form of SNRP. The problem has been formulated in a variety of

ways but, here, I will follow Karen Bennett in breaking it down as problem with a set of seemingly incompatible claims (2003):

1. The distinctness of the mental from the physical: *mental properties are not reducible to physical properties.*

Recall that, for Davidson, the mental is not reducible to the physical. One might be tempted to think that this commitment to the irreducibility of the mental implies that the mental is not physical at all. Indeed, the motto of the non-reductivist/token-physicalist – *"Mental events are physical events, but mental properties are not physical properties"* – implies a commitment to some kind of strong property dualism, where mental properties are non-physical. To this, however, the adherent to SNRP would reply that to say that the mental is not physical is to say that the mental is not reducible to the physical in a narrow sense; that is, mental kinds are not identified with explicitly physical kinds. So, for instance, say with mental event *e*, we have the property *P* of *being a belief*. This property P is not reducible to any explicitly physical kinds such as a brain state kind *B*. Mental kinds, for the SNRP, are physical in a broad sense; they are "higher-level" properties like *rigidity*[17].

2. The causal closure of the physical: *the physical is both causally closed and complete, so any effect e is wholly determined by a physical cause C (which we can specify in micro-physical terms).*

This claim denies that there can be any causes outside the realm of the physical, such as so-called emergent properties. To be committed to such properties is to deny that underlying,

[17] Properties like *rigidity* are "higher-level" in the sense that they are not found at the micro-level.

(explicitly) physical causes are sufficient to bring about certain effects. This is wildly implausible.

3. The causal efficacy of the mental: *at least some mental events cause physical events in virtue of their mental properties.*

It seems correct (to most of us, at least) to say that, for instance, Mary's believing that Saul was a murderer, combined with her desire not to die, caused her to run, where running is an explicitly physical event, or one that we can describe without reference to mental states.

4. No overdetermination: *the effects of the mental are not systematically overdetermined.*

An effect is overdetermined if it has at least two sufficient causes. Consider, for instance, a paradigmatic case of overdetermination by a firing squad. When someone is killed by firing squad, it is not necessary for all guns to be shot for the person to be killed. Imagine a case where we have two gunmen and one person to be executed. Both guns are shot and the person dies. In this case, the shooting of each respective bullet is a sufficient condition for the person's death; it is not necessary to have both. So, in this case, the effect – the person's death – is overdetermined, as it has two sufficient causes. It is unlikely that the effects of the mental are like the firing range case.

5. Causal exclusion: *if an effect has a physical cause, it cannot have a mental cause unless it is overdetermined.*

Claim 5 implies that if a mental event *e* causes an effect *e**, the physical properties *P* of e will exclude the mental properties *M* from being involved in the causal relation.

All five above claims have prima facie appeal, yet they seem to be incompatible. Following Kim, we can see that if a given mental event e has mental properties *M* and physical properties *P*, then 1 in conjunction with 2 and 3 commit us to holding that the effects of the mental have two sufficient causes: M and P. But if this is right, then, given 5, the effects of the mental are overdetermined. But 4 commits us to holding that they can't be overdetermined; hence the tension between these five claims. In light of this incompatibility, we have to reject at least one of the claims. It would be difficult to reject 2, 4, or 5, since rejecting any of these would have drastic implications for how we conceive of the world. We could reject 3, but that would lead to epiphenomenalism. We could reject 1, as Kim does (or at least attempts to do), and this would mean that we must be reductionists with respect to the mental. What to do?

§ 1.21: *One Proposed Solution: Functional Reduction*

For Kim, in order to avoid the problem of mental causes being excluded by physical causes, the most likely candidate for rejection is 1. That is, we must reject the claim that mental properties are distinct from physical properties. I agree with Kim that we should reject 1, but, as we shall see, his account actually entails the acceptance of 1; and so it fails.

As a functionalist, Kim thinks that mental kinds are multiply realizable. So, for instance, a mental kind *M* might be instantiated by a potentially infinite number of physical kinds *P1*, *P2*, *P3*, etc. In terms of mental causation, his picture of the relationship between the mental and the physical is as follows. Imagine that Bill's belief that Sally is single, in conjunction with his desire to date her, causes him to ask her out on a date. In this case, we

have a mental property *M1* (his belief), *M2* (his desire), and physical property *P* (the behavior of asking her out on a date). For Kim, the conjunction of M1 and M2 causes P in virtue of the physical properties realizing M1 and M2; indeed, this must be the case, otherwise we would have to reject 2 from above (the causal closure of the physical). The physical properties realizing M1 and M2 are the *supervenient bases* P1 and P2.

In light of this hypothetical case from above, the causal exclusion problem becomes evident. If we can give a complete causal account of P by appealing only to how it is caused by the conjunction of P1 and P2, then what work is there left for M1 and M2 to do? If we accept 1-5 from earlier, then the mental is excluded by the physical. For Kim, this is the fundamental problem for SNRP.

Kim's suggestion for resolving the causal exclusion problem is rejecting 1. This means we have to commit ourselves to the reduction of mental properties to physical properties. What it means to reduce a property *x* to a property *y*, however, is not uncontroversial, as there are at least a couple of views concerning this. First, it is traditionally thought that reduction requires biconditional bridge laws linking the xs and ys; this is known as *Nagelian reduction* (Kim 1998, pg. 26). So, for instance, we might see a correlation between pain and c-fiber firings. If it is true that 'pain' and 'c-fiber firings' are coextensional, then we can say that there are pains if and only if there are c-fiber firings. These bridge laws allow us to see how a "higher-level" theory such as psychology might be related to a "lower-level" theory such as neuroscience. The problem with Nagelian reduction, as Kim rightly notes, is that such biconditional claims are consistent with a variety of incompatible views,

such as substance dualism and emergentism (2005, pg. 22). While this problem might not be the same as the supervenience problem from above (since bi-conditional relationships are weaker than supervenience), the problem with it is similar: Nagelian bridge laws don't tell us much about how x is related to y. Further, for Kim, the multiple realizability of the mental bars such reductions, since no individual mental kinds are coextensive with any physical kinds.

In order to fix the problems noted with construing "reduction" as a biconditional relationship, a slightly modified form of Nagelian reductionism might construe 'reduction' to mean something like theoretical identification. So, for instance, water reduces to H2O in the sense that "they" are identical. For Kim, multiple realizability is just as much a problem with this view as it is with the Nagelian view. Further, Kim holds that identity statements don't explain anything as they simply "rewrite the rules" in a physical vocabulary for what has already been explained in a folk vocabulary (2005, pg. 145).

Given that he rejects the viability of the two aforementioned views for reduction, Kim opts for what he calls *functional reduction*. To functionally reduce a property is to define it in terms of its causal role. For example, consider the property of *being dormitive* (having the propensity to cause drowsiness). This property is multiply realizable and, as such, we cannot reduce it to any single kind of physical property. For Kim, it is a "higher-level" property; specifically, it is a second-order property, or the property of having another property. In the case of dormitivity, 'dormitivity' may be functionally defined as the property of having the first-order property of causing drowsiness (where the first-order property is the

basic, explicitly physical property). So, any object *o* has the property of *being dormitive* if one of its explicitly physical properties has a propensity to cause drowsiness. To give another example of functional reduction (one that concerns the mental), we might functionally define 'pain' as anything that plays the intermediary role between typical inputs such as tissue damage and typical outputs such as the normal types of resulting behavioral manifestations.

For Kim, the instantiation of the mental property *M* of *being in pain* at time 1 is nothing over and above the instantiation of the particular realizer *P* at that time (2005, pg. 26). One might think that to say that a particular instance of M is "nothing over and above" the particular instance of P is to say that M is identical to P, but Kim cannot be committed to this. Mental properties, after all, are second-order properties, while the properties on which they supervene are first-order properties. A given second-order property cannot be identical to the first-order property in question. As Ned Block observes, a second-order property is, by definition, not a first-order property (forthcoming, pg. 6). So, if John is in pain, there is the instantiation of the supervenient base property *P* – a first-order property – and the instantiation of the mental property *M* – a second-order property. Kim is, at the end of the day, a property dualist.

If Kim is a property dualist with respect to the mental and the physical, then he has not actually rejected claim 1 – the idea that mental properties are not reducible to physical properties – from earlier, as he is committed to holding that mental properties are distinct from physical properties. For him, the relationship between the mental and physical seems to

be that of constitution. With respect to pain, for instance, he holds that pain and its realizers are "intimately related" (2005, pg. 26). This view runs into a couple of problems. First, on this view, mental properties are nevertheless excluded by the physical properties, as the properties of the physical realizers are doing the causal work. Kim's response to this worry is to offer up what he calls the *causal-inheritance principle* – the idea that the "higher-level", supervenient properties "inherit" the causal powers of the "lower-level" base properties. Disregarding the *ad hoc* nature of positing such a principle, Kim's account still runs into the problem of overdetermination. I will consider his response to this problem in the next section, which concerns compatibilist attempts to circumvent overdetermination.

Even if Kim is able to assuage our worries about overdetermination, his view fails to give us an account of qualia – our ultimate concern here. This problem stems from the possibility of an inverted spectrum of qualitative mental states (Shoemaker 1982, Block 1990). Focusing on visual qualia for the moment, let us imagine we have two persons: Nonvert and Invert. Nonvert's color spectrum is typical for the population. When he sees an object *o* that everyone considers to have the color red, he experiences the qualitative sensation of red, or a token of the type of sensation any other person would have. Invert, on the other hand, is a qualia invert – at least with respect to visual qualia. He agrees that o is red, but the corresponding sensation in his mind is actually what normal persons would consider green, had they access to his experiences. Further, his entire color spectrum is completely opposite that of anyone else's. Despite this difference, there are no behavioral differences between

Nonvert and Invert. Indeed, nobody, including Invert, himself, would know that he is a qualia invert.

To qualify the above thought experiment, we should be clear about the fact that the lack of a behavioral difference, itself, does not establish the falsity of functionalism (Block and Fodor 1972). Functionalism is the idea that mental states are constituted by their functional role in the entire cognitive system. So, for the functionalist, it is possible that there may be a set of twins who are behaviorally indistinguishable, yet whose minds function differently (and, so, have different mental states). To be more precise about functionalism, then, we might say that the doctrine's fundamental commitment is the following supervenience thesis: *there can be no mental difference without a functional difference*. In terms of our thought experiment, let us say that Invert's case shows this supervenience thesis to be false; had Invert grown up with normal visual qualia, it would have made no difference to the internal (functional) workings of his mind.

The possibility of an inverted qualia spectrum poses a serious problem for functionalism. If we try to functionalize the experience of redness, we would define it in terms of its causal role as an intermediary between typical inputs and outputs. So, whatever plays that role is, by definition, red. The typical inputs and outputs for Nonvert and Invert are the same, so, according to our account, they are both experiencing red. But we know that Invert is not experiencing red, so functionalism cannot give us an account of what redness is.

Kim, himself, recognizes that inverted spectra are a possibility and, thus, his account fails to capture qualia. This failure is why he titled his latest book Physicalism, or Something

Near Enough, as he acknowledges that, yes, functionalism fails at accounting for qualia, but it does, in his view, give us an account of all non-qualitative mental states such as propositional attitudes – and this is as close to physicalism with respect to the mental as we will get (2005). Others, as it might be expected, are not so eager to accept the defeat of functionalism.

Some have objected to the very possibility of an inverted spectrum on verificationist grounds (Dennett 1991, pgs. 310-311). Those coming from this angle argue that there is no possible way to empirically verify that someone might be a qualia invert, so it is not meaningful to posit such a possibility. Verificationism, of course, it not taken very seriously these days, as it seems like a clear case of confusing metaphysics with epistemology. Problems with verificationism aside, I am not convinced that the possibility of a qualia invert is not empirically verifiable, in the first place. Recall that in the previous chapter, we discussed the possibility of cable neurons connecting one part of a brain to another. In principle, we could do the same thing with everyone on the planet. For example, a researcher could shut off their primary visual cortex and hook it up to someone else's.

Another objection to the possibility of an inverted spectrum comes from the idea that qualitative states may have an accompanying affective component (Campbell 2000). Consider that it is common for different colors to elicit different feelings in people. For instance, blue might make someone feel calm, while yellow might make them feel uneasy. These resulting states of calmness or uneasiness manifest themselves behaviorally, so it is argued that a qualia invert will also have inverted affective states. For example, Jill might

enter a room colored purple (it is considered purple by everyone else) and start to feel uneasy, since she is actually experiencing yellowness. The problem with such an appeal to a connection between qualitative states and affective states, however, is that unless it can be established that any given affective state is associated with a given qualitative mental state with metaphysical necessity (and I don't see how it can be), it is possible for there to be a qualia invert whose accompanying affective states are not inverted. All it takes is one possible case to show that an inverted spectrum is possible; and this is enough to be a problem for functionalism. There may very well be actual cases of individuals with the conjunction of inverted qualia and inverted affective states, but this does not help the functionalist.

Finally, one might object to our inverted spectrum argument by questioning how we are justified in claiming that inverted spectra are metaphysically possible in the first place. We might be tempted to reply by arguing the following:

1. Inverted spectra are conceivable.
2. Anything that is conceivable is possible
3. So, inverted spectra are possible.

This argument will do us no good, however, since premise 2 is false.[18] For instance, it is conceivable that water may have turned out to be something other than H2O, but given that it *is* H2O, it cannot be anything other than H2O, since identity holds with necessity.[19] In this case, the fact that we can conceive that water might have been something other than H2O only establishes that such a possibility is of the epistemic variety. Metaphysically

[18] This is not uncontroversial. I will defend this assumption in more detail in the 4th chapter.
[19] For a good discussion of epistemic possibility, see (Soames 2006, pgs. 196-199).

speaking, it is not possible that water is anything other than H2O. So, we cannot use the aforementioned argument as a strategy.

If conceivability doesn't entail possibility, another strategy we might try is establishing the nomological possibility of an inverted spectrum, since anything nomologically possible is metaphysically possible. Establishing the nomological possibility of an inverted spectrum, however, has obvious difficulties, since it would be behaviorally and functionally undetectable. Instead, we might try establishing the nomological possibility of a different kind of inversion scenario that establishes the falsity of the supervenience thesis the functionalist is committed to. Recall that the functionalist must hold that there can be no mental differences if there are no functional differences. Ned Block's inverted earth scenario establishes that this supervenience thesis is false by inverting the environment instead of the qualitative color spectrum (1990). The thought experiment goes as follows. Imagine that you, an Earthling, are fitted with "color inverting lenses" and are then whisked away to Inverted Earth, where the colors of the objects in the world are complementary to those on Earth. With the lenses on, however, you notice no qualitative difference. Further, the meanings of the color words the people on Inverted Earth employ are inverted; that is, they match up with your experiences. So, for instance, when mentioning the color of a fire hydrant, an Inverted Earthling would use the word 'red', even though she is experiencing green. You, on the hand, experience red when looking at the fire hydrant. For the functionalist, a mental kind is defined in terms of its typical inputs and outputs. In this case,

we have an instance of the qualitative experience of red with two distinct inputs – Earthly and Inverted Earthly inputs.

As Block acknowledges, this thought experiment may be used in an argument against only one variety of functionalism – the kind that conceives of functional roles as "long-arm", or as including specific features of the external environment (pg. 58). To put it another way, the Inverted Earth case is a problem for functionalists who share Putnam's twin earth intuitions and are externalists about mental content (1975). The problem is that on Earth, when looking at a fire hydrant, your mental state M has the content c, which includes features of the fire hydrant, such as its spectral surface reflectance properties. On Inverted Earth, however, you are in state *M* with the content *c**. In terms of the supervenience thesis, then, we have a mental difference (the content) but no functional difference.

The Inverted Earth scenario is no problem, however, for "short-arm" functionalists who hold that inputs "start at the skin" (pg. 58). Short-arm functionalists are internalists about mental content, so, to use a variation of Putnam's phrase, they think mental content is all in the head (1975). In terms of inputs, the short-arm functionalist would hold that, in the case of the fire hydrant, the input starts with the eyes, not the fire hydrant. In terms of causal roles, the short-arm functionalist would hold that qualia are defined by the role they play in the mind.

Now, as Block notes, it is possible that one might adopt some sort of "two-factor" theory wherein it is held that non-qualitative states are defined by long-arm roles, while qualitative states are defined by their short-arm roles, but it is extremely difficult to see how

such an account might work (pg. 70). For such a proposal to work, it must be the case that particular qualitative states have typical roles in the first place. Consider the qualitative experience of redness *R*. Does R have a typical causal role in one's internal psychology? I can't see how it does. This, of course, doesn't establish that it doesn't but, as Block argues, it does seem to establish that the burden of proof is on the functionalist to show us how R might be defined in terms of its short-arm functional role.

To those who maintain a short-arm functionalist account may be worked out in time, as I shall argue in the next sub-section, any account that holds mental properties are distinct from physical properties runs into the exclusion problem, no matter how "tightly" the relationship between the two kinds of properties holds. Since the functionalist holds that mental (second-order) properties are distinct from their first-order realizers, they are in the same boat as the compatibilists. So, I reject what Block calls the "containment response" to the inverted spectrum objection, or the idea that functionalism might still work for non-qualitative states (pgs. 53-54).

On a final note, if a particular qualitative experience *R* has no functional role, one might worry that qualia, in general, have no functional role.[20] Luckily for us, however, this does not follow. On my account, while having the experience of a spectrum of colors is adaptive, the particular spectrum, itself, that one experiences is an exaptation (at least this is the case for color). To put it another way, qualia, in general, are adaptive and have a functional role, but it is not the case that all particular qualitative kinds have functional roles.

[20] Having a functional role is not the same thing as being *defined* by a functional role.

The problem with functionalism, then, is that it cannot account for fine-grained differences in qualitative states, where these differences have no *functional* reason for existing.

§ 1.22: *Another Proposed Solution: Compatibilism*

The doctrine of compatibilism is so named because of the commitment that holding the distinctness of the mental from the physical is compatible with the commitment that the mental is causally efficacious (notable defenses of this view include Horgan 2001, Bennett 2003). More specifically, in terms of our five claims from earlier, compatibilists reject causal exclusion (claim 5).

Given that compatibilists accept the distinctness of the mental from the physical (claim 1), the causal closure of the physical (claim 2) and the causal efficacy of the mental (claim 3), one might worry how they may hold that there is no systematic overdetermination (claim 4), since they admit that there are, indeed, two sufficient causes for any effect of a mental event. The strategy for resolving this tension goes as follows. First, we analyze our concept of *overdetermination* by focusing on paradigmatic cases. We might consider, for instance, what it is about a firing squad case that makes the effect overdetermined. Upon reflection, we see that if either one (but not both) of the riflemen had failed to shoot, the person being shot would have still been killed. So, what we have is the following conjunction of conditional claims: *if rifleman a had shot his gun without rifleman b shooting, the person would have died* AND *if rifleman b had shot his gun without rifleman a shooting, the person would have died*. In this case, we have two *independent* causes for one effect, and, for the compatibilist, that is what makes the effect overdetermined.

The second step for establishing the causal efficacy of the mental, despite its distinctness from the physical, involves demonstrating how cases of mental causation for the non-reductive physicalist are markedly different from paradigmatic cases of overdetermination. The picture for the compatibilist is as follows. We have mental event *e1* with physical properties *P* and mental properties *M*. This event e1, in virtue of properties P and M, causes event *e2* which has physical property *P**. Yet, P and M are not related like the two causes in our firing squad case, since they are metaphysically dependent upon one another, due to the fact that M supervenes on P. So, in cases of mental causation, we do have two sufficient causes, but they are *dependent*, where the dependence relation comes in virtue of the fact that M is materially constituted by P, but not identical to P (given the lack of type/type identities). The intuitive idea here is the relationship between these two causes is "tight enough" to think of them as one cause and, thus, it can assuage our worries about overdetermination (Bennett 2008, pg. 8). Of course, the effects of the mental are still arguably overdetermined in some sense, since we still have two sufficient causes. To respond to this worry, the compatibilist makes a distinction between vicious (or worrisome) and non-vicious (or non-worrisome) overdetermination. The idea is that having two metaphysically dependent causes is a case of the latter kind of overdetermination, and holding this is compatible with accepting the claim that the effects of the mental are not systematically overdetermined; so, for the compatibilist, in claim 4, we should construe 'overdetermined' in the sense that we have two independent causes.

We might question the compatibilist's attempt to hold that vicious overdetermination only occurs when we have two independent causes. As Alyssa Ney notes, if we accept that physicalism is a contingent thesis, we accept that there is a possible world in which we have two ghosts – *G1* and *G2* – who, as matter of metaphysical necessity, occupy the same spatiotemporal region (2007, pg. 490). For instance, wherever G1 goes, G2 goes, and however G1 contorts his ghostly body, G2's body likewise contorts. Further, we can imagine that G1 and G2 simultaneously cause someone to be frightened. In this case, there is a strong metaphysical dependency relationship between G1 and G2, but, intuitively, the frightening event is overdetermined.[21] What this means is that an effect may be overdetermined by two dependent causes. So, the compatibilist is wrong to hold that overdetermination only occurs when we have two independent causes.

The compatibilist is relying too much on the semantics of 'overdetermination'. Even if we grant that overdetermination requires two independent causes, we may say "Very well, there is no vicious overdetermination involved in mental causation, but a complete physical cause, nevertheless, makes a mental cause *redundant*, if the former is distinct from the latter." This consideration puts us right back where we started. So, let us grant for the moment that the issue is about redundancy and not overdetermination. One way to eliminate this redundancy would be to hold that the mental is identical to the physical; this is a route the

[21] For those skeptical of an appeal to ghosts, Ney mentions the case of a boy who must bring something yellow or round to class for show and tell. To satisfy the criteria, he brings a tennis ball to class, which is both yellow and round. In this case, these two properties – the tennis ball's *being yellow* and *being round* – overdetermine the effect, despite being present in the same spatiotemporal region.

non-reductive physicalist obviously cannot take. Are there any other ways to eliminate this redundancy, given that we have rejected an appeal to the constitution relation? The problem is that the non-reductive physicalist needs a way to hold that the mental and the physical may be thought of as one cause, despite being numerically distinct. How can two distinct sets of properties be thought of as one cause? I confess, it is beyond my powers of imagination to envision x and y as one thing if they are not identical.

In sum, the non-reductive physicalist cannot deny that physical causes make mental causes redundant.

§ 1.3: *The Explanatory Exclusion Problem*

In response to our discussion of compatibilism, Barry Loewer has responded that those appealing to the exclusion problem are operating with a misguided theory of causation and, as such, the objection against SNRP does not get off the ground (Loewer 2002, pg. 659). That is, we are thinking about causation as *production*, the idea that when x causes y, it does so in virtue of a transferal of powers from x to y. According to Loewer, causation is not a fundamental physical notion (2002, pg. 661). When we look at classical mechanics at the microphysical level, what we have, in essence, are point particles governed by fundamental laws. That is, the laws direct the to-ing and fro-ing of the particles; it does not come in virtue of the properties of the particles, themselves. Instead, according to Loewer, in light of our well-confirmed understanding of the micro-world, we should think of causation as counterfactual dependence. If causation simply is counterfactual dependence, then the non-reductive physicalist has no problem, as it would follow that mental properties are causally

efficacious in virtue of being part of a metaphysically dependent relationship with the physical.

If we should construe 'causation' as counterfactual dependence (and I am not sure we should), then on our formulation of the exclusion problem the proponent of SNRP is free to hold that the mental is distinct from the physical. There is, however, another way to formulate the exclusion problem which shows us that the proponent of SNRP is nevertheless in big trouble. I propose that, instead of thinking of the exclusion problem in terms of causation, we might do better to think of it in terms of *explanation*. That is, if I am right, a complete physical explanation of the world excludes a robust explanation of the mental on its own terms.

Two camps have been approaching this same problem from different angles. On the one hand, we have Kim thinking about physicalism with respect to the mental in terms of causation. On the other hand, we have Fodor thinking about this issue in terms of explanations and laws. Fodor holds that an explanation of the nature of the mental will come only from the special sciences – namely, *psychology* for the mental (1997). Following Loewer (2009), there is a problem, as Fodor's account runs into an explanatory overdetermination problem. To see this, we might make some changes to our original formulation of the exclusion problem[22] by omitting references to causation and putting it in terms of explanation. Doing this, we get what might be called the *explanatory exclusion problem*:

[22] Pgs. 38-40.

1. The autonomy of mental explanation from physical explanation: *we can fully explain the nature of the mental without appealing to how the underlying (explicitly) physical properties (e.g. neurological properties) work*. For the NRP like Fodor we can fully explain the nature of the mental from a "higher-level" account, such as psychology. So, for instance, let's say we are given the following fact to explain: John ran away from Bob. Why did this happen? For the NRP, we explain this by appealing to, say, the fact that John believed that Bob wanted to harm him and, given that he is averse to being harmed, he believed that running away would satisfy his desire to remain unharmed.
2. The explanatory completeness of the physical: *for any physical state S1 at time T1, we can predict the nature of any subsequent state S2 at T2 with a complete set of physical laws*. The idea here is that not only are all physical facts completely physically determined, but an entity like Laplace's demon can have an *explanation* of all physical facts in terms of physics.
3. The nomological character of the mental/mental realism: *mental kinds are real in a scientifically respectable way, as they are governed by laws*. The idea here is that mental states are part of the fabric of the world in the same way that biological kinds like DNA or physical kinds like H2O are.
4. No explanatory overdetermination: *mental events are not (completely) explained by multiple sets of laws at a multiple explanatory "levels"*.

5. Explanatory exclusion: *if an event has a physical explanation, it cannot have a mental explanation unless it is explanatorily overdetermined.*

These five claims all have prima facie appeal. The problem, however, is that if we accept 2, it seems that an explanation of any given mental event is overdetermined if we accept that there are autonomous explanations coming from a special science like psychology. If this is right, it would be explanatorily superfluous to posit a separate and distinct existence of the mental. To put it another way, if physics can explain everything, why should there be a special science like psychology to explain the mental when it is already explained by physics?

The adherent to NRP like Fodor would respond to this problem by rejecting 2. For him, it is not that we have "higher-level" explanations of mental facts in addition to "lower-level" explanations; we only have "higher-level" explanations. Let us return to the case of John running away from Bob. For Fodor, this psychological explanation cannot be reduced in a way that appeals solely to the underlying (explicitly) physical properties for two reasons. Firstly, we have certain multiple realizability considerations. So, let's say that in our case John's desire to be unharmed is realized by the neurological property *N1*. Given that his desire might have been realized by a distinct neurological property *N2*, we cannot appeal to the neurological level to explain the fact in question, since there are no laws governing the disjunctive property [N1 or N2], and we need laws for explanations.

So, for Fodor, physics gives us laws that causally govern the mental, but not explanatorily; we must have another set of special science laws to explain the mental, and in virtue of the explanatory power of these laws, we posit the existence of distinct mental

properties. These laws, for Fodor, are non-strict and "ceteris paribus" (1991). Now, it certainly seems mysterious that there should exist fundamental laws other than the laws of physics to explain the mental. Arguably, it would be ideal if we could understand macro-level generalizations in terms of our understanding of the micro-level, given a preference for simplicity and parsimony. But, for Fodor, multiple realizability considerations force us into positing the existence of primitive special science laws. That is, the multiple realizability of macro-level properties bars the reduction of special science laws to fundamental physical laws. To put it another way, for Fodor and many other NRPs, intertheoretic reduction requires bridge laws, but multiple realizability considerations show us that such reductions cannot be had.

What if reduction, however, didn't require bridge laws? If this were true, then the multiple realizability of macro-level properties would no longer be a problem. One of the problems the NRP has been dealing with is that macro-level phenomena supervene on micro-level phenomena, but we seem to have no explanation for *how* the micro-level facts determine the macro-level facts. According to Loewer, the problem is that those like Fodor have only been seeing part of the picture. The laws of physics *themselves* don't determine the lawful regularities of the macro-world, as everyone agrees. For Loewer, however, the laws of physics in conjunction with initial conditions of the universe *do* determine the macro-level facts (2008, pg. 15).[23]

[23] For Loewer, we come to know the initial conditions from certain probabilistic constraints imposed by other background assumptions.

I will not work out the details of Loewer's account here since, for our purposes, giving a brief sketch should be sufficient to establish the plausibility of a set of more general and closely-related claims: (1) reducing a theory concerning macro-level phenomena to a theory concerning micro-level phenomena does not require bridge laws and (2) the multiple realizability of macro-level phenomena does not bar us from reduction. It is safe to say that claims 1 and 2 have generally been considered to be false by most philosophers after the multiple realizability arguments given by Putnam (1967) and the establishment of the purportedly autonomous discipline of cognitive science. As Loewer establishes, however, if we look at current practices in physics, we will see that there are other ways to think about reduction – ways other than Nagelian and functional reduction (see also Churchland 1985, Bickle 1997). In particular, David Albert has worked out a promising account concerning how we can derive the laws of thermodynamics from the classical dynamical laws of micro-physics with some modifications (Albert 2000). These accounts do not require bridge laws. Further, as Loewer notes, the phenomena of thermodynamics – like *temperature* – are multiply realizable (e.g. mean molecular motion in solids and mean molecular kinetic energy in gasses) like special science phenomena. Given that the phenomena of thermodynamics are multiply realizable and thermodynamics is reducible to classical mechanics, it follows that the multiple realizability of macro-level phenomena does not bar reduction.

The laws of thermodynamics, as Loewer notes, have more in common with special sciences laws than multiple realizability. Indeed, the similarities between the two are striking, and it is not a stretch to think that special science laws and the laws of thermodynamics are

alike in kind. If this is right, by analogy, we may argue that, since the laws of thermodynamics are reducible to classical dynamical laws with some modification it is reasonable to conclude that the special sciences are reducible to the laws governing the micro-physical world. Given that reduction doesn't require bridge laws and multiple realizability doesn't bar reduction, at least there is some room for such an account to be worked out. If this is right, then the principle of the explanatory completeness of the physical (claim 2) from the explanatory exclusion problem still seems promising. As such, the most plausible candidate for rejection or modification in the explanatory exclusion problem is the claim (1) that our understanding of the mental is autonomous from our understanding of the physical.

If we reject claim 1 but accept claim 2, then our understanding of the mental will come from physics. If Loewer is right, the regularities of the macro-world are physically determined by the goings-on of the micro-world. As such, the common objection that while Laplace's demon knows *that* a set of point particles x causes another set y to do z with lawful regularity, he does not know *why*, no longer holds. This is because, as Loewer notes, Laplace's demon has an understanding of how the macro and micro-world are related.

§ 2: A WEAKER FORM OF NRP

If what I have argued in the previous section is correct then, we may, in principle, explain the nature of the mental in explicitly physical terms. From this, one might assume that the special sciences have no role to play in our search to understand psychology and other macro-level phenomena. Luckily for the special scientists, as I shall argue, this is not

the case, as the special sciences explain macro-level phenomena in a different sense than physics does. In this section, I shall sketch out a distinction between these two senses of 'explain'. I shall then show how we might distinguish a weaker – and more viable – form of non-reductive physicalism from the stronger form discussed in the previous section. Finally, I shall end by arguing that it is better to understand Davidson's non-reductive account in this weaker sense, contrary to the general construal in the literature (and how we have so far construed his account).

§ 2.1: *Two Senses of 'Explain'*

Why did the chicken cross the road? The answer to this question is obvious to anyone with complete knowledge of the universe and infinite processing power: *at time T1 we had a set of point-particles in state S1 evolving in accordance with the laws of physics to yield S2 at T2*. While such an explanation might satisfy Laplace's demon, the rest of us are inclined to respond that the chicken crossed the road to get to the other side. Or to put it another way, stipulating for the moment that chickens have intentional states, we might say that the chicken desired to be on the other side of the road and believed that crossing would satisfy this desire. Such an explanation is more intelligible to us than the former explanation. It is in light of such considerations that philosophers such as Davidson have generally construed 'explanation' as an intensional notion, as whether or not a statement about x counts as an explanation is dependent on the way in which it is described (2001). As we have seen in the previous sections, however, if it is correct that the physical world may be given a complete explanation at the level of physics, then statements like our "chicken crossing the road" one from above do, indeed, count as genuine explanations, but of a different sort.

As is the case with our chicken example, we have an event *e* with two descriptions: one at the level of the underlying physics, and one at the intentional level, and both of these count as explanations. Given that both descriptions count as explanations, we might think that we have a case of explanatory overdetermination. We can avoid this, however, by making the quite (independently) plausible distinction between two kinds of explanations: *fundamental* and *pragmatic*.[24] The former kind of explanation is extensional and comes only from physics or some other discipline that is plausibly reducible to physics, such as chemistry or biology. The plus side to these kinds of explanations is the presence of maximal internal coherence. All entities referred to in fundamental explanations are accounted for in terms of a few basic physical principles. In this way, fundamental explanations might be said to be more complete or robust than the latter. The down side to these kinds of explanations is that they are difficult to come by in practice – especially when we are concerned with wildly multiply realizable entities or properties. The latter kind of explanation is intensional and comes from either folk theories or the special sciences. The plus side to these kinds of explanations is their intelligibility, given our epistemic deficiencies. The down side, however, is that there will be less internal coherence than explanations in the former camp. For instance, if we explain why John ran away from Bob and include reference to certain intentional items such as *beliefs* and *desires*, we must treat[25] these entities as primitive. So,

[24] Block makes a similar distinction between opaque and transparent explanatory contexts, but only discusses it in terms of theoretical identifications of the type/type sort.
[25] 'Treat' being the operational word, since they are not ontologically primitive.

when we ask, for instance, what makes desires desirous, we have no explanation, lest we turn our pragmatic explanation into a fundamental one.

§ 2.2: *Weak Non-Reductive Physicalism*

Context will determine whether a fundamental or a pragmatic explanation will be best, so we should not be eliminativists with respect to folk theories or special science theories. In this way, the special sciences will and should enjoy a significant level of autonomy: an explanatory one, but not an ontological one. So, while it might be the case that, in principle, we can explain certain mental events at the fundamental level, a psychological description will still be most practical. Further, if we are ever going to see how the micro-level facts determine the macro-level facts, we must first know the macro-level facts, themselves. With respect to the special science of most interest to us, in order to reduce psychology, we must first *do* psychology. So, even if we are aiming for a fundamental explanation, the special sciences have an integral part to play. In light of this, we may still rightfully be non-reductive physicalists with respect to the special sciences, in general, and psychology, in particular. To put it another way, we may be non-reductive physicalists in the (weak) sense that pragmatic psychological explanations will not reduce to fundamental explanations while still retaining their practical explanatory value.

With this distinction in hand, we may reexamine our two exclusion problems from the previous section. Let us start with the explanatory exclusion problem. Recall that, given the plausible account of the existence of a complete physical explanation of the mental, we run into an overdetermination problem if we hold that there are "higher-level" explanations as well. We can dissolve this problem by construing 'explanation' in two different ways in the

first commitment ("The autonomy of mental explanation from physical explanation"). If we construe the first mention of 'explanation' as *pragmatic explanation* and construe the second mention of 'explanation' as *fundamental explanation*, there is no longer an inconsistency. In this sense, we can unproblematically say that there is a "higher-level" and a "lower-level" explanation of the mental.

Let us now return to the causal exclusion problem. Here, things get a bit trickier. The causal exclusion problem is not a problem for the WNRP, if we take a linguistic turn. Recall that according to SNRP, mental properties are not identified with first-order or "lower-level" properties. Rather, mental properties are "higher-level" properties. Following John Heil, it is my contention that if we eschew talk of properties and instead talk in terms of predicates, WNRP can avoid the problem of causal overdetermination, while maintaining that the mental is not reducible to the physical in some sense (Heil 2003). So, if we do this, our first commitment ("The distinctness of the mental from the physical") in the causal exclusion problem should be interpreted in the following way: *descriptions of mental events using a mentalistic vocabulary (i.e. using mental predicates) are not reducible to descriptions using a physicalistic vocabulary (i.e. using physical predicates)*. Now, following Davidson, we only have laws governing mental events in a physicalistic vocabulary. But this is no problem because, given that we are not talking about properties, it does not follow that the mental is not causally efficacious; it only follows that when we describe mental events with a mentalistic vocabulary, we obscure what is going on nomologically and, hence, causally.

One might object to our appeal to predicates in the causal exclusion problem by arguing that we have unwittingly led ourselves into eliminativism. If, for instance, 'desire' neither picks out a distinct "higher-level" property (given the problems with causal efficacy and explanatory overdetermination), nor a distinct "lower-level" property (given multiple realizability considerations), then we might seem to be forced into a deflationary account wherein we must interpret uses of terms such as 'desire' in a mentalistic vocabulary as parts of vague predicates whose instances pick out individual properties that seem to have nothing in common with one another.[26] Such a deflationary account appears eliminativistic. However, we can supplement this account with the following sketch of an account of property individuation. Consider the heterogeneous disjunction *[P1 or P2]*. Let us stipulate that P1 and P2 are (nomologically, speaking) the only properties that may "realize" the property of *being a desire*. Let us further stipulate that these two properties are dissimilar in some sense (e.g. P1 is biological while P2 is silicon). Why not simply identify the property of *being a desire* with the disjunctive property [P1 or P2]? Some have argued against the possibility of such disjunctive properties on the grounds that they cannot be causally efficacious, since we have no laws with predicates containing references to disjunctive properties (Fodor 1997). But, given our account, we can consistently hold the following set of claims: (1) individual property instances are causally efficacious because they are governed by the fundamental laws of physics; and (2) each property instance is a token of the type *being a desire* in virtue of the particular roles they play in macro-level (counterfactually supported) generalities which,

[26] This is Heil's account.

recall, are explained by physics. Finally, given that the set of nomologically possible realizers does not exhaust the set of metaphysically possible realizers (at least for non-qualitative mental states), we should identify the property of *being a desire* with the disjunctive property that includes all metaphysically possible realizers.[27]

In light of our discussion of WNRP, we are now in position to return to Davidson's account. It is my contention that it is best to interpret his account as a commitment to WNRP and not SNRP, contrary to how most have construed him (McLaughlin 1992, Kim 2003). As I stated earlier, it is natural to formulate the doctrine of token-physicalism in terms of properties, where we are token-physicalists with respect to mental event *e*, just in case we hold that e has two distinct properties: mental and physical. If I am right, however, this formulation is incorrect. If we pay attention to Davidson's original formulation of token-physicalism, we can see that he is speaking in terms of descriptions and predicates, and this is not something we should gloss over (1970). For him, the mental is not reducible to the physical in the sense that *descriptions* of mental events using a mentalistic vocabulary are not equivalent to *descriptions* of those same events using a physicalistic vocabulary. There are plausible reasons that he might hold this. Firstly, for him, as we have seen with our discussion of intensional explanations, purely physical descriptions of mental events don't explain the event in question. This might seem inconsistent with our account, but recall, for Davidson explanations are intensional, and so he is using 'explanation' in the *pragmatic*

[27] The nitty gritty details of this account are not *that* crucial for our purposes. Here, I am just trying to make room for some disjunctive property account of non-qualitative properties.

sense, which is perfectly consistent with the claim that mental events can be explained in physical terms, if we interpret this use of 'explain' in the *fundamental* sense.

Secondly, for Davidson, mental events are not governed by laws, *so described*. That is, there are no laws containing mental predicates. Given that he thinks of laws as linguistic, it does not follow that he must be committed to the idea that there are no laws governing mental properties. Recall that on our account the laws of physics govern mental properties, but these laws do not contain mental predicates. These properties are physical *and* mental, despite there being no psychophysical laws. Davidson's account is perfectly consistent with this view; indeed, it isn't a stretch to think this *is* his view. If this is right, it shouldn't be too surprising, given his methodological commitment to a minimal ontology.

§ 2.3: *Where Qualia Fit Into Our Picture*

Let us return to qualia – our ultimate concern. Non-qualitative and qualitative mental states have an obvious difference, as the former are intentional while the latter are, well, qualitative. As such, it is relatively common for philosophers to treat them differently. Smart, for instance, held that we can account for intentional states in behavioristic terms, while he argued that we should be identity theorists with respect to qualia.[28] Somewhat similarly, Block has previously argued – with his "containment response" to the inverted earth problem – that functionalism will nevertheless work for intentional states, while we should look elsewhere to account for qualia. Kim, too, treats qualia differently, though in a more pessimistic manner, as he concludes that qualia are epiphenomenal. What these

[28] Of course, he thought of the identities as contingent.

accounts have in common is the commitment to the idea that we must have markedly distinctive accounts for the two types of mental states.

In some views, intentional states and qualitative states don't just differ in terms of the explanatory accounts we must have in order to understand their natures, they are distinct in a metaphysically deep sense. For instance, as Block has recently argued, Kim's functionalist account of the mental is not physicalist, despite Kim's insistence; it is a distinct kind of functionalist ontology. If we consider that if functional properties are not reducible to physical properties, then that means that functional properties *are not* physical properties, despite the fact that the realizers of the functional properties are physical properties. So, if we ask the functionalist the innocent question "What *are* mental states?", the functionalist must respond that mental states are functional things – end of story.

For Davidson, qualia also get a different kind of treatment, as there is a conspicuous lack of any mention of qualitative mental states in his writings (we might say that qualia get the silent treatment from him). From this, it might be easy to lump his account in with other accounts treating qualia in a (metaphysically) special way. I'm inclined to say, however, that his account is not at odds with the existence of qualia, in principle. We can only speculate, but perhaps we find no mention of qualia because he didn't see why we should posit the existence of qualitative properties in order to explain human behavior. If what I have argued in the previous chapter is right, however, there is a place for qualia after all in our explanatory picture of behavior. So, while Davidson was right to argue that we must posit the existence of beliefs and desires for explanatory purposes, he was either wrong to discount

the explanatory role qualia play or he had yet to figure out how to fit them into his overall project. I'm inclined to think the latter is true, given the progressive nature of his work. My account, then, is a supplement to Davidson's, as it adds (in a weak sense) sensations to the ontological inventory of psychology.

If I am right to maintain that qualia play the same kind of theoretical role as intentional states do in our Davidsonian project, then it follows that (1) we may also explain them in a pragmatic way and (2) WNRP is an account of the mental, in general. One might object that this can't be quite right, though, because with some "higher-level" WNRP account such as functionalism we can explain the ultimate nature of intentional states, but not qualia. But functionalism – the most plausible "higher-level" metaphysical account of the mental yet to be offered – is not only wrong with respect to qualia, but for intentional states, as well. What we have, instead, is the "higher-level" special science of psychology as a pragmatic explanatory account of the mental, which posits the existence of sensations, along with beliefs and desires.

While psychology explains mental phenomena in its own proprietary way, when it comes to the theoretical entities themselves, we have no further explanation expressible in the language of psychology itself. So, as we stated before, *beliefs*, *desires*, and (now) *sensations* are unexplained primitives. Now, we can and do give pragmatic explanations of these entities in the language of folk psychology. For instance, we might try to analyze our ordinary concept DESIRE in a way that appeals to other concepts such as WANTS. Likewise, when we are trying to analyze our concept PAIN, we might appeal to related concepts such as HURTS.

But, at the end of the day, these explanations are circular, and so the language of folk psychology, too, is insufficient to explain what they are in a robust or fundamental sense. To put it another way, we may elucidate the natures of these entities in question in a given language *L*, but we cannot *fully* define them within L, as long as L is the language of a special or folk science.

Sociologically speaking, elucidating the natures of intentional states in the above way seems satisfactory to us for the most part. Even though we may only fully explain the nature of intentional phenomena in physical terms, intentional states come so freely that it doesn't bother us. For instance, even though we have yet to construct artificially intelligent creatures with intentional states, it is not so difficult for us to imagine their existence – and it is reasonable to expect that they will exist in practice someday. Qualitative states, on the other hand, don't come so freely; and so it is a mystery to us why we should be conscious, while other intelligent life forms might not be. Further, qualitative states just *seem* like metaphysically different kinds of things than intentional states altogether. On our account, we can grant that qualia are, indeed, different kinds of states than intentional ones, and thus account for our intuition concerning their difference; it just so happens that the difference is physical. In terms of our argumentative strategy, this means that the intuitive thrust behind certain arguments against physicalism with respect to qualia is taken down a notch. The kind of arguments I have in mind are those operating under the assumption that, while we can easily account for intentional states in physicalist terms, qualitative states are mysteriously elusive (see Chalmers 1996). If I am right, and all mental states are in the same boat, this

means that we can give the following intuition pump: *the problem of qualia is no more a problem than the problem of reconciling the existence of the mental, in general, in the physical world; but, since the mental just has to be physical, qualia, too, must be physical.*

Now, insisting that qualia *just have to be physical* is not likely to completely assuage our deep-seated, dualistic intuitions. Luckily for us, the fact that qualia are type identical to physical properties means that, unlike intentional properties, we can have a fundamental explanation of their natures *in practice*. That is, we may have a reductive explanatory account of qualia in practice; we will discuss this in detail in the next chapter. We might liken the problem of qualia, then, to the problem of understanding how other kinds of macro-level properties such as *liquidity* can arise in a seemingly emergent way from their constitutive chemical properties; we will discuss this in detail in the final chapter.

CONCLUSION

In Section 1, we looked at Davidson's argument for NRP. Construing his account as a version of SNRP, we evaluated it in light of the causal exclusion problem and found that it implies that mental properties are causally inefficacious. We then examined Kim's alternative – functionalism – to Davidson's account of the mental and found that it, too, runs into the causal exclusion problem. As an account of qualia, specifically, functionalism has an additional problem, as the possibility of inverted qualia implies that qualitative mental properties cannot be defined in terms of their functional role. Finally, we discussed compatibilism as a way to get past the causal exclusion problem. Other problems aside, as we have seen, this account runs into an *explanatory* exclusion problem.

In Section 2, we turned from SNRP to WNRP. As I argued, given the multiple failures of SNRP, WNRP is the only viable form of NRP, as it properly distinguishes two senses of 'explain' – one fundamental and one pragmatic. For WNRP, to say that we cannot explain the nature of the mental in physical terms is simply to say that we cannot intelligibly or pragmatically do so, while acknowledging that there is some underlying fundamental physical account, at the end of the day. From this discussion of WNRP, we then looked back to Davidson's initial argument for NRP. As I argued, given that Davidson thinks of laws as linguistic and his talk of the mental is really talk of mental predicates, it is best to interpret his account as a form of WNRP. So, it looks like Davidson was right about the mental, all along.

CHAPTER THREE
THE IDENTITY THEORY

As we have seen from the previous chapters, if we are realists about qualia and think they play a causal/explanatory role in the world, then we should accept that they are physical properties in a reductive sense. In this chapter I shall argue that the reductive account that we should accept is *the (type) identity theory*[29], the idea that qualitative mental kinds like *pain* are identifiable with physical kinds like *c-fiber firings*.

Though we can find fragments of the identity theory earlier, the theory as we understand it does not make a full appearance until the late 1950s with J.J.C. Smart's landmark paper "Sensations and Brain Processes" (1959). For this reason, much of this chapter will pivot around Smart's initial construal and discussion. It is important to note that, while I am in complete agreement with Smart with respect to the core thesis he defends – the idea that any given kind of qualitative mental state *is* a kind of neurological state – at certain points, when appropriate, I shall either supplement or diverge from his account when it comes to establishing why we should accept the identity theory.

The structure of the chapter is as follows. In Section 1, I shall focus primarily on Smart's arguments for the identity theory, since, despite recent interest in the identity theory, there is little contemporary discussion concerning how best to interpret his commitments. In Section 2, I shall discuss and respond to some more recent objections to the identity theory.

[29] As opposed to the token identity theory mentioned in § 1.1.

§ 1: THE IDENTITY THEORY AND ITS HISTORY

For the most part, 'the identity theory' carries a negative connotation in the minds of contemporary philosophers. Like other widely rejected doctrines such as verificationism and phenomenalism, it has been filed away, to be retrieved only when historical curiosity strikes. If my thesis is correct, however, we should look to the identity theory not just because we are curious, but because it is highly plausible, despite its (hasty and ultimately unfounded) rejection. Indeed, given that non-reductive token-identity theories, as we have seen, are ultimately inadequate for helping us understand (in a strong sense of 'understand') how the mind fits into the physical world, the identity theory should look all the more appealing as a viable candidate.

§ 1.1: *SMART'S MODERN PREDECESSORS*

Incarnations of the identity theory can be found as far back as the 1930s (Carnap 1932, Schlick 1935), but it wasn't until the 1950s that it gained a significant foothold in the philosophical landscape with the works of Herbert Feigl (1958), Ullin Place (1956), and, most importantly, J.J.C. Smart (1959).

Feigl can be credited with providing a significant underlying motivation behind the theory, as it was he who coined the term 'nomological danglers', or the idea that it would just be strange if sensations were irreducibly psychical entities because that would mean it would be impossible to capture them with the nomological net of physical theory. To put it another way, the intuition is that if qualia are nomological danglers, then we would have a complete, mechanistic explanation of all the features of the universe except for qualia. For reasons having to do with ontological and explanatory parsimony, this just seems unlikely.

From here, we can credit Place for moving the theory forward by arguing that qualitative mental states *are* physical states.

Place's view, however, differs from ours, as his construal of 'are' or 'is' (singular) is that of constitution. For instance, to say that a statue *S* is a lump *L* is not to hold that S and L are one and the same, but simply to hold that L materially constitutes S in the sense that they are spatially coincident but numerically distinct entities. Given that in the previous chapter we have rejected appeals to constitution in this sense, Place's identity theory is not our identity theory.

§ 1.2: *Smart's Positive Account*

It is not until Smart's important work that we get the *identity theory* as we are construing it. As such, the rest of this section will revolve around Smart's discussion. Let us start with some context. Behaviorism was in vogue in the 1950s and with this came the commitment that all of the features of the world could be explained, in principle, by physics. Along these lines, it was thought that if the mental world is part of the physical world, we should then be able to reduce talk of the mental to talk of something at least implicitly physical such as overt behavior and dispositions. At this time, Gilbert Ryle's (1949) view was particularly influential. For Ryle, to say, for instance, that George wants ice cream is simply to say something like "he is disposed to yell 'yes!' when asked if he wants to go to Dairy Queen". For Smart, Ryle's account seemed right for a good portion of the mental (propositional attitudes), but not for all of it. When it came to qualitative mental states like sensations it seemed incorrect to say, for instance, that being in pain was nothing over and above something like being disposed to groan; in this case, the mental facts just seem

underdetermined by the behavioral facts. So, for Smart, while such dispositions may very well be correlated with sensations, they are not, themselves, sensations.

Influenced by Feigl and Place, Smart thought that – contra the Rylean view – it might be better to identify sensations with brain processes. Unlike Place's account, to say that a sensation *is* a brain process is to identify sensations with brain processes in the "strict sense" of identity. For him, sensations are not merely constituted by brain processes; they are "nothing over and above" brain processes, just as lightning is nothing over and above electrical discharge. So, pains are not simply *correlated* with c-fiber firings, as nothing can be correlated with itself; pains just *are* (something like[30]) c-fiber firings. At the time philosophers thought that identity statements must express propositions that are knowable a priori. Given that it is not obvious to us, however, that pains are c-fiber firings (if they are brain processes at all), the statement "pains are c-fiber firings" certainly does not *seem*[31] to express a proposition that we can know a priori. So, these identities seem to be knowable only a posteriori. As such, it seemed to follow that these identity statements were only contingently true. That is, though it may be the case that sensations are brain processes as a matter of fact, Smart concedes that it is logically possible[32] that they aren't. So, unlike

[30] It is important to note that it is not crucial that he be right about the *c-fiber firings* aspect in order for his account to be tenable. Pains are simply whatever a fully worked out neuroscience tells us they are. The important thing is that they are identifiable with a kind of physical state. In this light, it has been customary to think of use of 'c-fiber firings' as a place-holder for whatever pains actually turn out to be.

[31] Later in this chapter, I shall defend the claim that such propositions are nevertheless knowable a priori.

[32] Smart doesn't distinguish logical from metaphysical possibility here. But this would not be uncommon for the time.

identity statements that are necessarily true and knowable a priori like "All triangles are trilaterals", the identity statements in question are akin to what Smart called *scientific identities,* such as "Water is H2O" which he thought were only contingently true.

As Saul Kripke has shown (1980), Smart is wrong to hold that identity statements like "pains are c-fiber firings" express contingent propositions. We won't get into the details for why this is the case, but it should be obvious that identity is a necessary relationship an entity has with itself, since nothing could possibly fail to be itself. So, for our purposes, we shall diverge from Smart and consider the identity statements in question to express propositions that are necessarily true, if true at all.

Now, as it turns out, our paradigmatic claim of psychoneural identity – *c-fiber firings* and *pains* – is false, as the firing of c-fibers is only one element of the neurophysiology of pain (Hardcastle 1997). Despite this, philosophers (including myself) continue to use 'c-fiber firings' as a place holder for whatever neuroscience tells us is perfectly correlated with pains. Given that, strictly speaking, we have rejected the identification of c-fiber firings with pains, one might object that we have little reason to think that kinds of qualitative mental states are correlated with kinds of brain states, in the first place. Luckily for the identity theorist, there are, in fact, such correlations. Indeed, there is a whole research program (Crick and Koch 1990) in neuroscience that concerns itself with discovering the neural correlates of consciousness (NCC). For example, as Ned Block and Robert Stalnaker note, research for NCC suggests that instances of kinds of visual qualia are perfectly correlated with certain activity in the primary visual cortex (1999).

Let us assume that the neuroscientists are correct and there are, indeed, psychoneural correlations. Following Christopher Hill (1991), let us call the claim that there are such correlations *the correlation thesis* (CT). The truth of CT obviously doesn't imply that the correlated states are identical, but we can use the claim as a premise in what Hill – echoing Smart's appeal to Occam's Razor – calls the *best explanation argument*. The idea is that, if CT is true, then we need an explanation for why it is true. For example, if it is a fact that pains are perfectly correlated with c-fiber firings, then we need[33] an explanation for this fact.

To explain CT, the non-reductionists have two options. The first option is to hold that – as the epiphenomenalist or the emergentist might hold – there are, in addition to basic laws of physics, primitive psychophysical laws that causally link the mental to the physical. The second option for the non-reductionist is to hold that the relationship between the mental and physical is not causally necessitated but metaphysically necessitated. The idea is that mental properties supervene on physical properties, and that this supervenience is explained by something like constitution.

As we have seen from before, we have good reason for rejecting both non-reductionist accounts. Apart from these reasons, these two options fail in another respect, as they don't give us the best explanation for CT. Consider that, for the reductionist (construing 'reduction' in terms of identity), what explains CT is simply the identification of the mental states with the brain states. If this is correct, then, first, we don't need to posit the

[33] Those who aren't bothered by primitive facts might object that we don't *need* an explanation. I think we would all agree, however, that it is certainly preferable to have explanations.

existence of a set of laws in addition to the laws of physics. In this respect, the identity theory is a better explanation, as it is simpler. Second, if this is correct, then we have an explanation for why the mental supervenes on the physical; we needn't resign ourselves to holding that this relationship is primitive.

One might object to our explanatory account here by wondering how identity can explain anything in the first place. After all, on the face of it, it certainly seems that saying, for instance, that A is identical to A is explanatorily vacuous. For now, following Hill, I shall just give a brief sketch of how identification might serve as explanatory, as we shall come back to this issue in more detail in the next section. Consider, for instance, that wherever Superman is, Clark Kent is. That is, the spatiotemporal location of Superman is perfectly correlated with the spatiotemporal location of Clark Kent. Observing this, Lois Lane might wonder *why* this is the case. That is, she might wonder why the statement "Clark Kent is present if and only if Superman is present" is true. If our analogy to the identity theory holds up, then we can explain to Lois Lane why this statement is true by appealing to the fact that Superman just *is* Clark Kent. Intuitively, this seems correct. When Lois Lane learns that Superman is Clark Kent, she has an explanation for why Clark Kent is always around where Superman is.

§ 1.3: *A Negative Account*

As we saw before, incarnations of the identity theory were around decades before the account currently in question. Arguably, these earlier incarnations didn't have any significant sway in the philosophical community because of a certain set of objections which might be grouped together by the fact that they are primarily semantic or epistemic. Smart addresses

these objections in his larger attempt to show that the identity theory was too hastily rejected at the time.

§ 1.31: *Semantic/Epistemic Objections*

The general structure of the semantic/epistemic objections is the following: by Leibniz's Law, if *S* (a sensation) is *B* (a brain process), then S and B share all their properties; S has a property that B doesn't (or vice versa); so, it follows that S is not B. For instance, it seems to be true that I can know that I am in pain at a given time, but, at the same time, not know that my c-fibers are firing. So, by Leibniz's Law, it seems to follow that pains aren't c-fiber firings.

Smart's reply is something to the effect of the following: *I can know that a lightning strike can kill a person but not know that an electrical discharge can kill a person; this doesn't mean that lightning isn't electrical discharge. We have to discover that lightning is electrical discharge empirically* (1959, pg. 152). The problem with this objection, at root, is that 'know' in this case determines an intensional context, while the conclusion concerns the extension of the terms 'pain' and 'c-fiber firings'.

Like the epistemic objections, the semantic objections don't pay heed to the crucial distinction between intension and extension. One such objection is the following: 'pain' doesn't mean the same thing as 'c-fiber firings', so pain aren't c-fiber firings. To this, Smart has an obvious reply, which is similar to his reply to the epistemic objections: 'the morning star' and 'the evening star' don't have to mean the same thing in order for the entities referred to by those terms to be the same.

I'm inclined to say that Smart's responses to the aforementioned epistemic/semantic objections are fine enough as they stand, since, as mentioned earlier, it is a truism that we cannot derive ontological/metaphysical conclusions from semantic/epistemic premises. Now, there are more sophisticated contemporary arguments against the identity theory in the literature that rely on epistemic/semantic premises, but we shall wait for the next chapter to discuss them, as they are first and foremost arguments against physicalism, in general.

§ 1.32: *Metaphysical Objections*

Let us now consider another set of objections to the identity theory: those concerning primarily metaphysical issues. To many of these objections, as we shall see, Smart has a response. Smart's responses, however, have failed to sway the contemporary philosophical community in the way that his responses to the previous set of objections did during his lifetime. So, here, at times, I shall diverge from him and supply what I take to be better responses to these objections.

Smart relies heavily on drawing an analogy to cases like "the morning star" and "the evening star". Along these lines, however, as Smart notes, one might object that the reason that we have distinct concepts corresponding to these distinct descriptions of the same object is because we have two distinct properties (1959, pg. 148). That is, the object to which we are referring – Venus – can be described in these two ways because it has two properties: *being the morning star* and *being the evening star*. If this is right, then it seems that it is in virtue of the purported fact that a brain state has two distinct properties that we are able to refer to it in two different ways: mental properties and physical properties. For instance, a

given brain state might have the property of *being a c-fiber firing* and the property of *being a pain*. The former property is physical, while the second is mental, so it seems to follow that we are nevertheless committed to a form of dualism in the sense that we have two ontologically distinct kinds of properties, instead of two distinct kinds of substances. Smart replies to this objection by arguing that we do not, in fact, identify the referent of 'sensations' with the referent of (certain) 'brain processes' by the mental property *sensations*; rather, we identify them in a way that is "topic-neutral", or neutral with respect to the ontological status of what is being identified (1959, pg. 150). For instance, the property of, say, *having a yellowish after-image* is identified, not by special mental properties, but by the typical causes that bring it about, like a particular kind of reflectance property[34]. Construing mental properties in this way is to construe them in a quasi-functional way. That is, a given mental property is characterized in terms of its role within a certain cause/effect relationship[35]. As such, it does not follow that whatever plays this role is either distinctly mental or distinctly physical.

There is a problem with Smart's response to this objection, however, as the only way we might be able to fully characterize a mental state in this way is if we go the functionalist route and include its causal relationship within its interaction, not only with the world, but with other mental states. But, as we have seen in the previous chapter, functionalism doesn't work. Though it might be the case that we may give *some sort* of functional characterization

[34] Smart does not characterize colors in this way in this paper, but he does later.
[35] Later, Armstrong and Lewis give a more fully developed functional account of mental properties that defines mental properties in terms of their causal role between other mental states and the world.

of the mental, such a characterization certainly won't work with qualia, given the inverted qualia objection; qualia just aren't functionalizable.

Another route we might try is to question whether the fact that we have two modes of presentation for x means that these two modes of presentation must correspond to numerically distinct properties. This seems obviously wrong. Consider that we might refer to a sample of water with both the predicate "is an instance of H2O" and "is an instance of water." In this case, we have two different modes of presentation expressed by these two predicates. Even if we grant that each mode of presentation must correspond to some property, it doesn't follow that these properties must be numerically distinct. For instance, we might have the "micro-level" property *C* (*being a sample H2O*) and the "macro-level" property *W* (*being a sample of water*). Given that water just *is* H2O, C and W are identical. Likewise, in the case of pain, for event *e*, we have two different modes of presentation corresponding to the following properties: *M* (*pain*) and *P* (*c-fiber firings*). As in the case with water, it doesn't follow that M and P are numerically distinct. So, contra Smart, this objection doesn't force us into some quasi-functional topical neutral account, as the multiplicity of modes of presentation doesn't imply the multiplicity of distinct properties.

§ 1.321: *Kripke's Modal Objection*

Another metaphysical objection – of a modal nature – to the identity theory comes from Kripke (1980, pg. 148). For the most part, before Kripke's seminal work "Naming and Necessity", philosophers thought that identities could be contingent. For instance, it is a contingent fact that Aristotle was the teacher of Alexander. Despite the contingency of this

fact, it was held that we could nevertheless say that Aristotle is *identical* with the teacher of Alexander. Kripke showed us, however, that identity is a not a contingent relationship, but a necessary one. In the case of 'Aristotle' and 'the teacher of Alexander', the former term is what Kripke calls a *rigid-designator*, while the latter is *non-rigid*. That is, 'Aristotle' rigidly designates the individual Aristotle in the sense that our use of the name picks out the same individual in all possible worlds. 'The teacher of Alexander', on the other hand, is a non-rigid designator in the sense that there are worlds in which the expression picks out a different individual; for instance, we can imagine that another philosopher was the teacher of Alexander. To put it in terms of properties, to say that Aristotle is the teacher of Alexander is to say that *being the teacher of Alexander* is a contingent property of Aristotle's, as there is a world in which Aristotle is not the teacher of Alexander but a soldier.

Kripke's analysis of proper names like 'Aristotle' extends to general terms like 'heat'. For Kripke, 'heat', like 'Aristotle', is a rigid-designator – one that rigidly designates *mean molecular motion*. So, there are no possible worlds in which there is heat but no molecular motion. Now, one might reply that it certainly *seems* that we can conceive of heat without molecular motion, but as Kripke rightly argues, this "seemingness" is just that. When we think that we can conceive of heat without molecular motion, what we are really doing in confusing *heat* with another property that is only contingently associated with molecular motion: the *sensation of heat*. For example, when we think we are imagining a camp fire without heat, we are really only imagining a camp fire without the sensation of heat – a case where, say, there are no sentient beings in the vicinity. Despite the lack of the sensation of

heat, there is nevertheless that which generally causes the sensation of heat: heat. So, when we say that we cannot conceive of heat without molecular motion, what we mean is that we cannot *clearly* and *distinctly* conceive of heat without molecular motion. To put it another way, to say that x is conceivable is to say that x is conceivable in an ideal situation or with an idealized agent. In the case of identity, if we want to maintain that x = y, despite the intuitive appearance of a contingent relationship, the burden is on us to explain away this intuition – as has been done in the *heat/molecular motion* case – by showing that we cannot clearly and distinctly conceive of x without y.

With Kripke's terminology in place, let us now discuss his argument for the non-identity of *pain* with *c-fiber firings*. For Kripke, 'pain' is a rigid-designator like 'heat'. That is, when we think of all actual and counterfactual states of affairs, 'pain' always picks out the property of *being felt as pain*. So, there are no possible worlds in which a pain exists but is not also felt as a pain; this sounds plausible enough. Though pain may be materially constituted by c-fiber firings in this world, we can clearly and distinctly conceive of a world in which we have pains but not c-fiber firings. For instance, we can imagine Data the android being in pain with his positronic brain. Given that it is conceivable that pains exist without c-fiber firings, it is metaphysically possible that they are instantiated without c-fiber firings. As such, 'pain' does not rigidly designate *c-fiber firings*. Thus, it follows that pains are not identical to c-fiber firings. So, Kripke might agree with Smart that pains are contingently identical (in some sense) with c-fiber firings, but he would argue that this means they are actually not

identical in the first place, as the notion of *contingent identity* is entirely wrong-headed. We may formulate this argument in a more formal way with the following:

(1a) We can (clearly and distinctly) conceive of the instantiation of pains without the instantiation of c-fiber firings.

(2a) Whatever is conceivable is (metaphysically) possible.

(3a) If it is possible that x may be present without the presence of y, then x and y are not identical.

(4a) It is possible to have pains without c-fiber firings.

(5a) So, pains are not c-fiber firings.

This argument, though focusing on pains and c-fiber firings, may be formulated in a more general way if we replace 'pains' and 'c-fiber firings' with *sensations* and *brain processes*, respectively. Doing this, we get the conclusion that sensations, in general, are not brain processes. For illustrative purposes, however, we shall focus on the argument as it is stated (above).

Now, it is important to note that, if we can not only conceive of the presence of sensations in individuals with other kinds of physical states (e.g. positronic states) but non-physical bodies such as ghosts, as well, then Kripke's argument has a much stronger conclusion than "pains are not c-fiber firings" or "sensations are not brain processes"; it implies that pains are not physical states, at all. As such, Kripke's argument is an argument against *any* form of physicalism. We shall not go into this matter here, however, as we shall

examine this more far-reaching argument against physicalism in the next chapter. For the moment, we shall focus on the restricted construal (from above) against the identity theory.

§ 1.3211: *SMART'S REPLY TO KRIPKE*

Given that the work of Smart's on which we are focusing predates Kripke's reply, we obviously find no response to this argument in the text. To my knowledge, Smart does not respond to Kripke in any of his later works, except for a brief comment in his entry in the Stanford Encyclopedia of Philosophy.[36] There, in reply to Kripke's challenge that we cannot explain away the apparent contingency of the relationship between pain and c-fiber firings as we can in the case of heat and molecular motion, Smart states: "There is a sense in which the connection of sensations (sensings) and brain processes is only half contingent. A complete description of the brain state or process (including causes and effects of it) would imply the report of inner experience..." What he appears to be saying here is that it only *seems* the statement "pains are c-fiber firings" is contingently true because we have incomplete knowledge of the workings of the brain. That is, it appears that Smart wants to hold that an idealized agent, fully grasping the meaning of "pains" and the meaning of "c-fiber firings", will be unable to conceive of the falsity of "pains are c-fiber firings". Of course, given that we are not idealized agents, we are not in a position to explain away the apparent contingency. With respect to our argument from above[37], then, Smart wants to reject the claim that we can conceive of the instantiation of pains without the instantiation of c-fiber firings (1a) [and, thus, the claim that it is possible to have pains without c-fiber firings (4a)]. In this

[36] <http://plato.stanford.edu/entries/mind-identity/>
[37] Pg. 86.

light, though he does not make this comparison himself, it seems right to hold Smart might be best considered an a priori physicalist like Frank Jackson (2005) (see also Smart 2006). An a priori physicalist holds that an idealized agent with a complete description of all the explicitly physical facts (e.g., a description in the language of physics or some language obviously reducible to physics) will also know, a priori, all of the facts concerning qualitative mental states.

I'm inclined to agree with Smart and reject (1a), as well. Indeed, if we have interpreted Smart's above remarks correctly about whether or not the statement "pains are c-fiber firings" expresses a proposition that is knowable a priori, I'm also inclined to agree that there is a plausible case to be made that we can know these identities a priori (but this case will need to be worked out in much more detail than Smart's cursory remarks – more on this later). For now, we shall see what others have written in reply to Kripke's objection. In particular, we shall focus on responses coming from Christopher Hill and Scott Soames.

§ 1.3212: *Hill's Reply to Kripke*

In reply to Kripke, Hill (1997) argues that we can explain away the illusion of contingency in the pain/c-fiber case by appealing to a distinction that Thomas Nagel (1974) makes between *perceptual imagination* and *sympathetic imagination*. On the one hand, to imagine something perceptually is to put ourselves in a state we would be in if we were actually perceiving that very thing. For instance, if we are to imagine that a tree has fallen over, we put ourselves in the kind of mental state we would be in if we were observing this to actually be the case. On the other hand, to imagine something sympathetically is to put ourselves in the very state in question. For instance, if we sympathetically imagine the

sensation of "redness", we would put ourselves in that very state by conjuring up some sort of red image.

For Hill, when we are imagining the instantiation of a pain without the instantiation of c-fiber firings, what were are doing is "splicing together" two images from these two different types of imagination. So, we are sympathetically imagining the state *being in pain* by intentionally approximating what it is like to instantiate that very state, while we are perceptually imagining the lack of c-fiber firings by putting ourselves in the kind of mental state we would be in if we were actually observing this to be the case. What all of this means is that even though, in some sense, it is right to say that we can imagine pains without c-fiber firings, this occurs only with two senses of 'imagine'. In terms of the claim that we can conceive of the instantiation of pains without the instantiation of c-fiber firings (1a) in Kripke's argument, this means (assuming that we are roughly use 'conceive' and 'imagine' interchangeably here) we are using 'conceive' in "we can (clearly and distinctly) conceive of the instantiation of pains" in a different way than we are in our implicit usage of 'conceive' in "without the instantiation of c-fiber firings*"*. As such, we are equivocating. Finally, as Hill notes, it is unlikely that were are in a position to know when we are and are not perceiving a brain state in the first place, as brain states might plausibly be thought of as being on the *theoretical* side of the line delineating the distinction between "theory" and "observation".

§ 1.3213: *Soames' Reply to Kripke*

Hill's response gives us good reason to question Kripke's conceivability claim (1a), but making the case for why the appearance is an illusion is not enough to break the illusion itself; we are still left with our Cartesian intuitions. Following Soames (2005), we might

argue that we are in error to try to meet Kripke's challenge to explain away this illusion in the first place. As Soames notes, there are two routes to the necessary a posteriori in *Naming and Necessity*. As Soames argues, however, these two routes are inconsistent. To help us locate the first route, consider a statement that we might use to express an a posteriori necessary truth such as "This table was originally made out of wood, necessarily." If we do not know that the table is, in fact, made of wood, we can imagine that it might be made of other kinds of material such as plastic. Assuming the doctrine of origin essentialism is correct, we can know a priori that certain features concerning the origin of the table are essential to it. Determining what these features are, though, requires empirical investigation. In the case of the table, we know a priori that *if* the table is made of wood, it is necessarily made of wood; knowing that it is wood requires us to look at the world. We can see that in this case, even though we could clearly and distinctly conceive that the table was made of plastic, it was nevertheless *necessarily* made of wood. So, conceivability in this case does not determine metaphysical possibility but only epistemic possibility. In the case of pains and c-fiber firings, on this account, when we are conceiving that pains are something other than c-fiber firings we are, at best, establishing that this is epistemically possible. We need knowledge of the actual state of affairs in order to determine what is metaphysically possible or impossible for pain. If the property of *being a pain* turns out to be a neurological property *N*, on this account it is necessarily N, despite the ability to conceive otherwise.

If the aforementioned account is a legitimate route to the necessary a posteriori, then we are warranted in rejecting this second route that Kripke uses to argue against the identity

theory, as it relies on the truth of the claim that whatever is conceivable is (metaphysically) possible (2a). While it might be legitimate to regard this second route as failing to establish that there are a posteriori necessary truths, we should not reject it as a route to the necessary a priori. This is important for us because one might object to our *application* of this first route to the necessary a posteriori, by arguing the following[38]:

(1b) The kind *pain* is identical to the kind *c-fiber firings.*

(2b) One can know a priori of the kind *c-fiber firings* that all of its instances involve the firing of c-fibers (this claim is to be interpreted *de re*).

(3b) By Leibniz's law it follows that one can know a priori of the kind *pain* that all of its instances involve the firing of c-fibers.

This conclusion is admittedly counterintuitive, but the identity theorist is forced into it, given that she cannot reject (1b). We might try to reject (2b), but if c-fiber firings are natural kinds like H2O (and it seems like they are), and it is plausibly the case that we can know a priori of the kind *H2O* that all of its instances include one part hydrogen and two parts oxygen, then it seems that we have no good reason to think that, mutatis mutandis, the same does not go for c-fiber firings. Given that our claims in this argument are interpreted *de re*, our use of Leibniz's law is legitimate and the conclusion certainly seems to follow.

Now, as an aside, it is important to note that this argument does not mean that the above account (the first route) of the necessary a posteriori is unsound, in general. For instance, if I say that "pain is c-fiber firings" and I am using 'is' as the *is of predication*, then

[38] I'm influenced by Teresa Robertson's "A Puzzle About Kinds" (forthcoming), here.

we no longer get the inference from (2b) to (3b). In this way, we are not barred from holding that, say, truths about one's origin are instances of the necessary a posteriori. Despite Place's protestations, though, the identity theorist must use 'is' as the *is of identity*. As such, the conclusion (3b) here is warranted.

Given the exclusion arguments from the previous chapter, we have good reason to think that qualitative kinds like *pain* are identical to neurological kinds like *c-fiber firings*. Given the argument in the previous paragraph, we have good reason to think that the statement "pains are c-fiber firings" expresses a proposition that is knowable a priori. So, if we are type-physicalists with respect to qualia, we must be a priori physicalists. We shall deal with this issue in more detail in the next chapter, but for now consider that Laplace's super demon from the previous chapter might plausibly be said to know a priori of pains that they are c-fiber firings. Further, though we may know a priori of pains that they are c-fiber firings, it does not follow that we may know a priori of any *instance* of pain that it is a c-fiber firing. For now, we still have to address Kripke's claim that it is knowable a priori that pains are not c-fiber firings.

§ 1.3214: *A Thoroughly Externalist Approach*

How might we explain away our Cartesian intuitions effectively? Kripke is right to hold that we don't come to identify pain by a contingent property like we do in the case of heat and normally coinciding heat sensations. I think he is unjustified, however, in claiming whatever feels pain-like is a pain. Consider the following. Drawing from Putnam's Twin Earth scenario, imagine that we have two samples of two different liquids: *H2O* and *XYZ*, respectively. Imagine further that, despite their different chemical compositions, these two

liquids are qualitatively indistinguishable at the "macro" level of description. Corresponding to this "level" of description, in a folk vocabulary, we get the terms 'water' and 'schwater', respectively. In this case it seems clear that even though these two samples appear to be samples of the same kind of thing, water and schwater are nevertheless distinct. I think the same sort of considerations apply in our case with pain and c-fiber firings. So, let us tweak the H2O/XYZ case to fit our purposes. Doing this, we get the following scenario involving two sentient creatures: Harry the human and Mary the Martian. At a given time, Harry is in the state of *being in pain*. At the same time, Mary, who lives in the same universe as Harry, is in the state of *being in schpain*, a mental state with properties that are qualitatively indistinguishable from the properties of pain. Harry's pains are instantiated by c-fiber firings, while Mary's schpains are instantiated by x-fiber firings. The question I want to pose here is the following: *even though pains and schpains are qualitatively indistinguishable, does it follow that they are numerically identical*? Intuitively, the answer might seem to be 'yes', but aside from this, it certainly doesn't seem to follow that qualitative indistinguishability implies numerical identity; the water/schwater case is a testament to this.

I'm inclined to think that, despite its prima facie counterintuitiveness, there is philosophical utility in accepting that pains are distinct from schpains (and the general implication falling out of this claim).[39] In terms of the objection coming from Kripke, we may reply by saying that when we think we can conceive of pains without c-fiber firings, what we are really conceiving is a situation in which somebody is experiencing something

[39] This will be discussed in the next few sub-sections.

pain-like, but not pain. Now, this reply to Kripke doesn't amount to an argument for why pains are not schpains, but it does show that we are not forced to reject the identity theory because of such arguments.

In response to my claim that pains are not schpains, one might wonder why we shouldn't just regard the kind *pain* as disjunctive. After all, it is plausible that non-qualitative mental kinds like *beliefs* are disjunctive. To answer this question, we might do well to look at actual scientific practices. The concept of *heat* that we find in thermodynamics applies to multiple distinct physical kinds. Despite this, it is generally regarded that we nevertheless have a textbook case of the reduction of thermodynamics to statistical mechanics. To illustrate this, let us make things simple and say that it has been established that 'heat' may apply not only to *X* (e.g. *heat in a gas*) but to *Y* (e.g. *heat in a solid*) and *Z* (e.g. *heat in a plasma*), as well. Now, we might hold that, given that we have had such a successful reduction of thermodynamics to statistical mechanics, regarding *heat* as a disjunctive kind is no obstacle to having a reductionist account of *heat*. I'm inclined to say, however, that appealing to physics in this way doesn't give us a full answer, as our question is ultimately ontological. Though we might use the term 'heat' to refer to cases like X, Y, and Z, it doesn't follow that X, Y, or Z are, themselves, identifiable with heat. At best, what we have is the disjunctive predicate *is heat* that might correctly be used when talking about X, Y, or Z but does not, itself, pick out a single, natural kind. To accept all of this is not to say that these heat variants X, Y, and Z have nothing in common, as, despite their differences, we might nevertheless say that they all fall under a larger class H. So, we have the class H whose

members are instances of X, Y, and Z. Seeing these class relations demonstrates how we can say that X, Y, and Z are numerically distinct kinds of things (from each other), while, at the same time, are all identical with respect to their membership in H. In this light, we can agree that Kripke is right in holding that the property *heat* is just what we have come to know as heat first and foremost, where other properties resembling heat are really something like *heat** and *heat*** (or Y and Z). If all of this is right, we might use the predicate *is pain* to refer both to pains and schpains, but it doesn't follow that pains and schpains are the same kind of state.

Another objection that one might make against the claim that pains are not schpains is by arguing that our analogy to water breaks down because the content of the concept WATER is broad, while the content of the concept PAIN is narrow. Broad content is that which is out in the world (the thing being represented). Narrow content is, to draw from Putnam, that which is "in the head". For example, imagine that you see a black cat while walking to campus. At this time, you are instantiating a mental state *M* that is both broad and narrow. The broad content of M is simply the situation involving the case. The qualitative aspect of M is the image of the cat and its surroundings. The components of the narrow content include the subjective experience of *blackness*. As Putnam has shown, it is possible to have two individuals who are instantiating qualitatively identical states but, nevertheless, numerically distinct kinds of states. For example, imagine that we have two worlds *w1* and *w2*. The only difference between w1 and w2 is that in w1 the term 'water' refers to H2O, while in w2 'water' refers to HXY. Let us imagine that in w1 I am looking at

a pond filled with what I call 'water'. At this very same time and corresponding place in w2, my twin is looking at a pond filled with what he calls 'water'. Despite the fact that our mental states are qualitatively identical, Putnam makes the compelling case that we are in different mental states.

Aside from the representationalists, it is generally held that what goes for H2O and XYZ does not go for qualitative mental states. In some sense, I think this is right, as it certainly seems right to say that the qualitative aspects of mental states are just what are in the head. In another sense, however, it seems that the concept PAIN, for instance, is also broad. If the content of PAIN is what is in the head and what is in the head is just what is in the brain, then the content of the concept PAIN is in the brain. So, when we use 'pain' to refer to someone else's pains, the broad content is their c-fiber firings. When we use 'pain' to refer to our own pains, the broad content is the firing of our c-fibers. If this is all right, then the representationalists aren't the only ones who can hold that the content of a particular qualitative state is broad. Allowing for this shows that the identity theorist, too, can accept the externalist consensus about mental content; we do not have to claim that qualia are an exception. Indeed, given that our concepts concerning qualitative mental properties are, at the end of the day, concepts concerning physical properties, the identity theorist *should* be an externalist. Consider, for instance how strange and ad hoc it would be to hold an internalist view of neurological properties, but an externalist view of all other kinds of properties.

§ 1.322: *After-Images*

Let us turn to a final metaphysical argument against the identity theory. As Smart discusses, one might object to the identity theory with the following argument:

(1c) My after-image has no spatial-temporal location;

(2c) My brain processes do have a spatial-temporal location:

(3c) So, by Leibniz's law, my after-image is not a brain process, nor is it a physical object.

Smart replies to this objection by distinguishing the *objects* of experiential states (e.g., qualitative mental states) from the *experiential states*, themselves. For him, the problem with this objection is identifying the mental object – in this case, the after-image – with the mental state. He holds that there are, in fact, no after-images. Like the discarded philosophical notion of *sense data*, their ontological status is fictional or instrumental. That is, talk of after-images is not talk about actual things. We do, however, have the *experience* of an after-image, and this does have a spatial-temporal location. So, further, while the after-image might be considered to have the property of *being orangish*, the experience of this after-image is not, itself, orange; to hold otherwise is to commit what Place (1954) calls the *phenomenological fallacy* – the fallacy of thinking there is something like a theater of the mind in which images masquerade. Likewise, the property of *having a pain in my leg* might be considered to be located in the leg, itself, but Smart thinks we are wrong to locate the pain there. Now, for him, the pain isn't in the brain, either; only the *experience* of the pain is in the brain.

In some sense, Smart is right to hold that the objects or content of mental states are not the states, themselves, as long as we restrict ourselves to the intentional objects. Smart is wrong, however, to hold that the content of a qualitative state is not a component of the state itself. Though qualitative mental states *might* have broad content, as we have seen with our discussion of representationalism in the previous chapter, they necessarily have narrow content. For Smart, the image is not any part of the mental state. Even if we grant Smart that there is only *the experience of the image,* the experience, for him, does not include the sensation of blackness.

If Smart's view from above seems like eliminativism (with respect to qualia), that is because (arguably) it is. While this might seem like an incredible claim, since he is often credited with formulating a first possible physicalist solution to the "hard problem of consciousness", this interpretation is consistent with his other writings. With respect to color sensations, Smart holds that colors [subjectively construed] are not part of the furniture of the world (1961). Colors, for him, only exist in the world in the Lockean sense that they are "powers" to produce in us certain experiences that allow us to make certain discriminations[40]; our ordinary subjective construal of 'colors' yields no referent. Instead of trying to figure out how we might understand how sensations such as *the experience of blackness* can exist in the physical world, Smart's strategy is just to hold that we are in error to think that there are such things in the first place. So, instead of thinking of Smart as giving us an account for

[40] Later he adopts Hilbert's view that colors are reflectance properties; these are still objective properties "out in the world."

purporting to solve the hard problem of consciousness, it might be better to say that Smart's strategy is to deny that there is such a problem in the first place.

One might think that this eliminativist implication must be inadvertent on Smart's part, as eliminativism and reductionism are generally thought of as in stark contrast to one another. At the time, however, as John Bickle argues, eliminativism was once thought of as a close cousin to reductionism (2005). The idea was that the predicates and terms in ordinary language that we use to refer to the mental are just too confused and dualistic to refer to anything actually existing. Instead, we should *replace* these aspects of our folk vocabulary with more scientifically respectable terms and predicates. The replacement of certain components of our folk vocabulary first requires the elimination of what is already there. This form of "replacement" reduction, then, is a form of eliminativism. Indeed, as Bickle notes, eight years after he published SABP, Smart expresses his sympathies with those holding a more explicit eliminativism such as Feyerabend. He states:

> I am even doubtful now whether it is necessary to give a physicalist analysis of sensation reports. Paul Feyerabend may be right in his contention that common sense is inevitably dualistic, and that common sense introspective reports are couched in a framework of a dualistic conceptual scheme.... In view of Bradley's criticisms of my translational form of the identity thesis, I suspect that I shall have to go over to a more Feyerabendian position (1967).

Contra Smart's own interpretation of himself, I'm inclined to say that instead of "going over" to Feyerabend's position, it might be more correct to say that Smart was already implicitly committed to eliminativism. So, this "going over" is not so much a shift in his position as it is a shift in his thinking about what his own view entails.

I'm a qualia realist and, so, I'm inclined to say that Smart's response to this particular metaphysical objection is not one that we want to help ourselves to. Rejecting Smart's solution, however, does add a constraint to our account, as holding that a given qualitative state is a physical state means that it must, indeed, have a spatiotemporal location. So, in the cat example from before, on our account the image of the cat and its surroundings must have a spatio-temporal location. On the face of it, this claim is counterintuitive, but it does mesh with other intuitions we have. For example, it is neither uncommon nor uncontroversial to say "Right now, I have a throbbing pain *in* my head." When we say something like this, our intuitions at the time are physicalistic to some degree, as we are committed to holding that pains have a spatiotemporal location (*right now* and *in my head*).

§ 1.33: *The Multiple Realzability Objection*

A final objection we shall discuss here is one that is primarily empirical. This is the *multiple realizability* objection that originates with Putnam (1967). Arguably, this objection is the one that has had the biggest negative effect on adherence to the identity theory. We might formulate the argument in the following way:

(1d) If the identity theory is correct, then pains are type-identical to some physical kind such as c-fiber firings.

(2d) If pains may be instantiated by multiple kinds of properties, then they are not necessarily instantiated by any single kind of physical property (such as c-fiber firings).

(3d) Pains may be instantiated by multiple kinds of physical properties (maybe non-physical properties, as well).

(4d) So, they are not necessarily identical to any single kind.

(5d) Given that identity is a necessary relationship between an entity and itself, pains are not type-identical with any single kind of property such as c-fiber firings.

(6d) So, the identity theory is false.

The idea is that mental states, including qualitative mental states, are, *as matter of fact*, not type identical with any types of physical states. For example, even though we may find pains to be instantiated by c-fiber firings in humans, when we look at other creatures in the world such as octopi, we see that pains can be "realized" by a different kind of brain state – or *physico-chemical state,* to use Putnam's terminology. Since the viability of the identity theory is predicated on the idea that there exist at least some type/type identifications, such as the identification of the qualitative mental type *pain* with the physical type *c-fiber firings*, the purported fact that sensations are multiply realized (and thus pains are multiply realized) seems to show that the identity theory is false.

As was the case with Kripke's objection, since Putnam's paper was published eight years after Smart's, Smart has no response to the multiple realizability objection. We can, however, speculate what he would say by looking at his account of topic neutral translations and (again) his brief comment in his SEP entry. As stated before, Smart's account of topic neutrality seems to be an early form of functionalism. Indeed, David Lewis and David Armstrong built upon Smart's account of topic neutrality in their formulation of functionalism. They argue that, instead of thinking of functionalism as a competitor to the identity theory, it might be better to think of functionalism as a route to an identity theory.

Smart echoes this sentiment in his SEP entry by stating that functionalism and the identity theory are actually not so different. Instead, we might think of functionalism as an identity theory in itself, where mental kinds are identified with second-order functional properties. Recall, however, that functionalism doesn't work, so we must find another way to respond to this argument.

Though some might question a few of these premises, I'm inclined to reject (3d). When Putnam published his landmark piece, he took it as being obvious that animals like octopi could feel pains without the same kinds of brain states. The intuition behind this claim is that it would just be chauvinistic to think that only creatures with our kinds of brain states could feel sensations. Echoing objections from Hill (1991) and Polger (2008), we might say that even if we hold that octopi can't feel pains, we needn't be committed to thinking that they can't have sensations, at all. *If* octopi have different kinds of brain states than we do, the identity theorist may still hold that they are capable of instantiating qualitative mental states but just different kinds of states. So, perhaps, creatures with different underlying neurophysiology have sensations that we don't have. Indeed, we might say that it is chauvinistic to insist that they must have the exact same kinds of sensations as we do.

Further, as William Bechtel and Jennifer Mundale note (1999), Putnam's conception of a brain state doesn't mesh with how neuroscientists actually individuate brain states. Much of the success of neuroscience is due to the fact that we have found commonalities between species, despite other differences. For instance, imagine that we have a person *Barry*

and an octopus *Larry*. Let us stipulate that Barry and Larry are both in the same state *being in pain* at a particular time. When we examine their brains, we might find differences, but there are nevertheless important commonalities between the two. In this case, drawing from actual data from neuroscience, Barry and Larry share the same coarse-grained property. That there are such coarse-grained properties should be unsurprising; even the philosopher's favorite natural kind *water* is identified with the coarse-grained property that we might conceptually extract from individual water samples. Consider that if we compare a glass of tap water with a glass of distilled water, there will be noticeable differences between the two. Despite these differences, these samples still share the property of *having H2O*. For Bechtel and Mundale, the error in the multiple realizability argument is that proponents are individuating qualitative mental states coarsely, while individuating brain states finely. For example, intuitively, some qualitative differences between the state Barry is in and the state Larry is in don't imply that they are not instantiating the same mental type *pain*. Indeed, there are qualitative differences between states that one person might be in at different times (e.g. the pain of a paper cut versus the pain of burn), but these differences don't imply that these are different mental states. On the other hand, the intuitive error that Putnam commits is thinking that brain states are finely individuated (i.e., small changes mean different states). As Bechtel and Mundale argue, when we make sure that we individuate qualitative and brain states in the same way (either both coarsely or both finely), then the intuitive force behind the multiple realizability argument is undercut.

As Polger notes, the multiple realizability argument comes in three strains. These three strains can be seen by considering how we can interpret 'may' in the third premise. First, when we say that pains "may be instantiated" we might construe 'may' in an *actual* sense. That is, we might say that pains may be instantiated by different brain states because they are, in fact, instantiated by different brain states. As we have seen, however, the argument for actual multiple realizability, in light of what we have argued so far, is not as convincing as it once was. Another way of interpreting 'may' would be in terms of physical or nomological possibility. So, one might hold that, though pains are not, as a matter of fact, instantiated by different kinds of brain states, it is nevertheless physically possible that they may be. For instance, if a given mental state is functionalizable, then it is plausible that it might be instantiated by different kinds of physical states. But, qualitative states are not functionalizable, so in the case of pains we cannot appeal to such considerations to establish the nomological possibility of being instantiated by different physical states. Aside from this, I can't think of how we might establish that pains are, in fact, multiply realizable in the nomological sense. So, at this juncture, we might just say that it is an open question that we will return to shortly. Finally, we might construe 'may' is a metaphysical sense. So, we might say that, even if pains are not multiply realizable (actually or nomologically), it is nevertheless the case that there is a possible world in which pains are not, say, c-fiber firings; as such, pains are not necessarily c-fiber firings and, thus, not c-fiber firings, at all. Establishing the metaphysical possibility of this would rely on the conceivability of such a scenario (if conceivability establishes metaphysical possibility, at all). As we have seen earlier, however,

arguments for the metaphysical possibility of pains being instantiated by something other than something like c-fiber firings aren't convincing.

Returning to the claim that it is nomologically possible that pains are multiply realizable, let us consider the best-case scenario that the proponent of MR might appeal to in order to establish the truth of their claim. The thought experiment is similar to the one above including Harry and Mary. Recall that, in this scenario, Harry is in pain while Mary is in schpain. Even if we grant the proponent of MR that it is nomologically possible that a drastically different kind of brain state might instantiate a pain-like property like *schpain*, as we have seen, it doesn't follow that pains and schpains are the same type of property.

In order to establish that pains and schpains are the same type of property, the proponent of MR must establish that *being felt as a pain* is a sufficient condition for being in pain. I have yet to see an argument for such a claim. I speculate that one might try to establish this by arguing that the meaning of 'being in pain' just means *being in a pain-like state*. As such, it follows by definition that schpains are pains. If we are semantic externalists, however, this conclusion doesn't follow, as we would need to know the ontological facts in order to establish what the meaning of 'being in pain' is. For example, semantic externalists hold that we can be wrong about the meaning of 'water', while still using the term competently. It is not until we find out that our use of 'water' refers to H2O that we find out what the word means. If we are semantic internalists, on the other hand, then, according to the consensus of philosophers of language, we are committed to the wrong theory of meaning.

So, intuitions aside, we have little reason to think that being in a state that feels like pain means that we are, indeed, in pain. As such, we have little reason to think that pains are multiply realized. On the other hand, in light of our discussion of the exclusion problem in the previous chapter, we have good reason to think that qualitative mental properties like *pain* are individuated by the coarse-grained grained physical properties by which they are identified.

§ 2: CONTEMPORARY OBJECTIONS TO THE IDENTITY THEORY

In this section, I shall sketch out and respond to some contemporary objections to the identity theory. Through these responses, we shall see some important implications our account has (e.g. the implication that we can have a fundamental explanation of qualia in practice).

Recall that in the previous section it seemed correct to hold that the fact expressed by the statement "Clark Kent is present if and only if Superman is present" is explained by the fact that Clark Kent is Superman. Kim has recently objected against the identity theorist's appeal to identity in this way (2005, pg. 135). The idea behind Kim's objection is that since identity is a relationship between an object and itself, to say that Clark Kent is Superman is simply to say that a = a. But to say that a = a is to say something tautologous and, thus, explanatorily vacuous. Kim is not wrong to hold this, but I am inclined to say that he is missing part of the picture. We can grant him that "Clark Kent is Superman" expresses the same proposition as "Clark Kent is Clark Kent." Granting this does not bar us from acknowledging that there is pragmatic element in the former statement that is not present in

the latter. When Lois learns that Clark Kent is Superman, though she does not gain any *de re* knowledge, she learns something *de dicto* – namely, that 'Clark Kent' and 'Superman' refer to the same person. Further, she learns that two corresponding sets of descriptions (Clark Kent-ish descriptions and Superman-ish descriptions) are descriptions of the same entity. So, on an intensional level, "Clark Kent is Clark Kent" differs explanatorily from "Clark Kent is Superman." In terms of pains and c-fiber firings, we can grant that the statements "pains are c-fiber firings" and "pains are pains" express the same proposition while at the same time holding that the former statement gives us an explanation at an intensional or pragmatic level that the latter does not.

Such identifications allow us not only to have a pragmatic explanation of the nature of *pain*, but a fundamental explanation in practice, as well. To get the gist, by analogy, consider how we might give a fundamental explanation of the features of water. It is uncontroversial that we should have a complete physical explanation of the nature of H2O (e.g., how it behaves). Following Block and Stalnaker (1999), let us consider a fact for which we might want an explanation: *that H2O freezes.* A fundamental explanation of this fact will come when we are able to describe the freezing of H2O in terms of the basic laws and constituents of physics. It is important to note that having an explanation of H2O in this way is not an intensional affair, but an extensional one, as we are finding out *what it is in virtue of* that H2O freezes. Along these lines, then, it follows that since water is identified with H2O, we have a fundamental explanation of its features (this explanatory fact is transitive since we are talking *de re*). Likewise, when we fully understand the nature of c-fiber

firings, if pains are c-fiber firings, it follows that we will fully understand the nature of pain at a fundamental level. Given that *is a pain* is not a wildly disjunctive predicate, it is plausible that we will have a fundamental explanation of the nature of pain *in practice*.

Let us now turn to another contemporary objection to the identity theory that Polger anticipates and attempts to refute (2011, pgs. 30-31). The problem for the identity theorist, Polger argues, is that if we consider physicalism to be a contingent thesis, it follows that the identity theory is false. The argument for this claim is the following reductio:

(1e) Sensations are identical to brain processes in all possible worlds. (Identity theory)

(2e) Physicalism is contingent; there are some non-physical worlds containing non-physical sensations. (Contingent physicalism)

(3e) There are some worlds in which sensations are not identical to brain processes.

(4e) The identity theory is false.

The problem is that if physicalism is a contingent fact, then it is possible that there are non-physical sensations. But, such a possibility means that sensations are not necessarily physical and, thus, not physical, simpliciter. Polger suggests that this implication forces the identity theorist into holding that physicalism is a necessary truth. If this is right, then there are no possible worlds in which there are non-physical things. Polger admits that this is a strong claim, but he thinks that the identity theorist must bite the bullet here.

It seems right to hold that if physicalism is a necessary truth, then we have no worlds in which sensations are not physical processes. That is, the metaphysical necessity of physicalism is a sufficient condition for there being no possible worlds containing non-

physical sensations. I think Polger is wrong in thinking that being committed to contingent physicalism means that we must admit that there are worlds in which sensations are non-physical. It doesn't seem right to hold that a necessary condition for holding that sensations are brain processes is that physicalism is a metaphysically necessary truth.

Let us differentiate two senses of 'physicalism' here. In one sense, *physicalism* is the thesis that *everything* is physical; let us call this *general physicalism* (GP), since we are thinking of it as a general claim. General physicalism comes in two varieties: contingent and necessary. If we are contingent and general physicalists (CAGP), then we interpret 'everything' as ranging over just the actual world; as such, it is a contingent truth, if true at all. If we are necessary and general physicalists (NAGP), then we construe 'everything' as ranging over all possible worlds; as such, it is a necessary truth, if true at all. In another sense, however, *physicalism* is a claim about, not everything, but *some* things; let us call this *restricted physicalism* (RP). This way of construing 'physicalism' also comes in a contingent variety (CARP) and necessary variety (NARP). For example, we might say that, as a matter of fact, the (arguably) functional property *intelligence* is instantiated by physical things; so, in this sense, we are CARPs with respect to intelligence. That is, despite the fact that *intelligence* is instantiated only by physical things in our world, we might nevertheless hold that there are worlds in which non-physical creatures are intelligent. Kinds such as *water*, on the other hand, arguably, may not be anything other than the physical compound H2O. So, we are NARPs with respect to water. Being NARPs, we may nevertheless grant that there are non-physical worlds; these worlds just don't contain water.

I think the above distinction is the natural and correct way of talking about physicalism. To see this, consider that if Polger is correct, and we reformulate the argument from above in terms of 'water' and 'H2O', we get the conclusion that water is only physical if we are NAGPs. This doesn't seem right. It is fine to say that water is necessarily physical though the universe isn't. One might respond that the analogy to water (again) breaks down because, with water, we may grant that there are worlds in which watery stuff is present, but not water. So, for instance, let us say that there is a possible world *w1* in which non-physical, watery stuff is present. Since *being watery stuff* is not a sufficient condition for being water, we may hold that this is a world in which water is not present.

Unlike water, when it comes to sensations, on the other hand, it certainly seems that a mental state with the property of *being sensation-like* is, ipso facto, a sensation. So, imagine a world *w2* in which there is a ghostly being instantiating a mental state with the property of *being sensation-like*. In this case, it seems right to say that a mental state's having the property of *being sensation-like* is a sufficient condition for it being a sensation.[41] Granting this doesn't mean the identity theory (construed correctly) is false, however. If we are NARPs with respect to particular kinds of sensations, we can avoid this implication, while also avoiding the strong claim that everything is physical. If this is right, then we should construe 'sensations' in Polger's argument to refer to the set of sensations that we find in the actual world. Interpreting the first premise in this way means that our claim is that all of the kinds

[41] Now, one might think that holding this contradicts what we said earlier in our response to Kripke. Recall that we said that x's *being pain-like* is not a sufficient condition for x's *being a pain*. The difference here, however, is that 'pain' picks out a single property.

of sensations we find in the actual world are necessarily physical. So, the identity theory is just a claim about the kinds of sensations that we find in our world.

A final contemporary objection comes from Hill, a former identity theorist turned respresentationalist (2009). As Hill notes, it is a general rule that we can make an appearance/reality distinction for any given phenomenon. On a standard construal of *the identity theory*, qualia seem to be an exception to this, as *X's* being in a pain-like state necessitates that X is, ipso facto, in pain.

For Hill, the only way in which we might have an appearance/reality distinction with respect to qualia is if we become representationalists and hold that the content of any given mental state is just its representational content, where the representational content is always "out in the world". For the representationalist, a qualitative mental state represents an external physical state *as being a certain way,* where this "certain way" of representing is how this physical state appears to us (as opposed to how it really is).

Given the non-standard way in which we have construed *the identity theory*, however, contra Hill, we can do justice to the appearance/reality distinction. Recall that earlier we questioned whether being in a pain-like state implied that one was, in fact, in pain. For us, it is possible for two agents to be in qualitatively indistinguishable mental states, but numerically distinct kinds of states. For example, imagine that in case 1 in *w1* I am in pain, a state which (minimally)[42] supervenes on the brain state *c-fiber firings*. In *w2* (still part of case 1), my twin is in schpain, a pain-like state that (minimally) supervenes on the brain state *x-*

[42] I use 'minimally' here to mean that we are, of course, open to it being the case that the states are identical.

fiber firings. These two brain states are drastically different such that they do not share even any coarse-grained physical properties. Now, imagine that a team of neurosurgeons, unbeknownst to us, open up my brain and my twin's brain and switches these states around, so that the pain-like state I am in is produced by x-fiber firings and the pain-like state my twin is in is produced by c-fiber firings. After this procedure, we then wake up and both have pain-like states; this is case 2. According to the account that has been defended here, I think that I am in pain even though I am really in schpain (and vice versa with my twin). So, it is possible to be in a scenario – case 2 – that is epistemically indistinguishable from another – case 1 – and not be in the same qualitative mental state. As such, it *appears* to me that I am in pain when I am not. So, we can do justice to the appearance/reality distinction; the representationalist does not have a monopoly in this area.

CONCLUSION

In Section 1 of this chapter, we looked at the history of the identity theory starting with Feigl and Place, then focusing on Smart. Smart's positive account of the identity theory starts with the consideration that sensations such as pains can't be wholly defined by their dispositional role. The rest of his positive account is primarily methodological, relying on the Occamist assumption that theoretical simplicity implies that we should hope for a physical account of sensation. As I argued, all of this seems correct, though some of the details of Smart's own account must be modified in light of certain philosophical advances (e.g., the recognition of the necessity of identity).

Smart's negative account comprises responses to two sets of objections. The first set relies on semantic/epistemic premises in order to attempt to establish that pains just can't be c-fiber firings. Smart's responses to these objections, as we saw, are fine as they stand. The second set of objections is primarily metaphysical in nature. As I argued, Smart's responses are problematic at several junctures. Indeed, from a historical perspective, it is interesting to note that in these responses, we can see that Smart is actually committed to a form of eliminativism with respect to sensations. In his stead, I replied to these objections with a novel account that includes a commitment to a version of a priori physicalism and a denial of the widespread assumption that, as far as qualia are concerned, likeness implies identity.

In Section 2, I responded to three contemporary objections to the identity theory. I first demonstrated that we can appeal to identity for explanatory purposes. I then employed the rejection of the assumption that likeness implies identity in my response both to the objection that physicalism about qualia implies physicalism, in general, and to the objection that the identity theory cannot do justice to the appearance/reality distinction.

CHAPTER FOUR
OBJECTIONS TO PHYSICALIST ACCOUNTS OF QUALIA

Though I have provided strong reasons in the previous chapters for thinking that qualia must be physical properties, an intuitive problem remains, as it nevertheless *seems* that qualia just aren't the same kind of thing as other explicitly physical properties. That is, we are committed (at least implicitly) to a dualistic ontology: mental and physical. And, intuitions aside, there are some seemingly compelling arguments that provide independent reasons for thinking that qualia are, indeed, non-physical and, thus, support our dualistic intuitions. In this chapter, I shall demonstrate how these arguments fail. I shall then sketch out an account of how the physicalist might make some headway in explaining away these admittedly powerful dualistic intuitions.

The structure of the chapter is as follows. In Section 1, I shall address what is known as the *conceivability argument* against physical accounts of qualia. Very briefly, the claim is that we can conceive of a world physically identical to ours but lacking qualia and so this world is metaphysically possible; hence, qualia aren't physical. I shall focus my discussion in this section on arguments from David Chalmers, as he is the one who has argued most forcefully in this vein. In Section 2, I shall address what is known as the *knowledge argument* against physical accounts of qualia. Very briefly (again), the idea is that we can know all the physical facts of the world without knowing any facts concerning qualitative aspects of the mental and, so, qualia aren't physical. I shall focus my discussion in this section on an argument from Frank Jackson, as it is he who has set the terms of the current debate. Finally,

in Section 3 I shall sketch out the implications of our discussion on the problem of the so-called *explanatory gap* (Levine 1983), the problem concerning how "Technicolor phenomenology can arise from soggy grey matter" McGinn (1989).

§ 1: THE CONCEIVABILITY ARGUMENT

One might properly identify two conceivability arguments against physical accounts of qualia. The first, which comes from Kripke, we discussed in the previous chapter. Recall that, for Kripke, a particular state of affairs that is conceivable is also metaphysically possible as long we are careful in our conceiving not to confuse that state of affairs with another, separate state of affairs (e.g., not to conflate a situation involving the presence of heat with a situation involving just the presence of the sensation of heat).

Building on Kripke's initial insight concerning the relationship between conceivability and metaphysical possibility, Chalmers has constructed a relatively sophisticated framework known as *two-dimensional semantics* that, he argues, provides another route (separate from Kripke's) to metaphysical possibility from conceivability. In this section I shall outline the two-dimensional argument against physical accounts of qualia, then demonstrate how it fails.

§ 1.1: *TWO-DIMENSIONALISM AND CONCEIVABILITY*

Despite considerable differences between different physical (e.g., non-reductive and reductive) accounts of qualia, all accounts are committed to the claim that facts concerning qualitative mental properties supervene[43] on explicitly physical facts. More formally, let 'P' refer to all of the micro-physical properties in our world and let 'Q' refer to all of the

[43] I'm using 'supervene' here in a neutral way.

qualitative mental properties in our world. The physicalist with respect to qualia is committed to the following conditional claim *PST* (the physicalist's supervenience thesis): □(P → Q). To put it another way, we are committed to holding that a physical duplicate of our world necessarily duplicates the mental features of our world. So, there are no metaphysically possible worlds containing the micro-physical properties of our world but not containing the qualitative mental properties of our world [□~(P and ~Q)]. If there are, then all forms of physicalism with respect to qualia are in trouble. More specifically, on the one hand the ontologically reductive physicalist is in trouble, since if *being a pain* is a physical property, it must be physical in all possible worlds. For the ontologically non-reductive physicalist, on the other hand, the violation of supervenience means that physical properties don't have much to do with qualitative mental properties.

Chalmers denies the truth of PST (1996, 2003), holding that qualia aren't physical and, therefore, we are wrong to look to the physical world for an explanation of what qualia are. For him, we cannot solve the *hard problem of consciousness* with any physical account. If he is right, then our project – and any project like it – for trying to understand qualia is hopeless. We might formulate Chalmers' argument in the following way[44]:

(1a) If the doctrine of physicalism with respect to qualia is correct, then, minimally, PST is true.

(2a) Anything ideally (i.e. we aren't confused or lacking relevant information) conceivable is metaphysically possible.

[44] This looks a lot like Kripke's argument, though Kripke doesn't talk about physicalism, in general.

(3a) We can conceive of a world in which we have P but not Q (from PST). (For example, we can conceive of a world with the same micro-physical properties as ours but lacking qualia. This world might be called a *zombie world*, since our counterparts in this world look and act like us but lack qualia.)

(4a) Zombie worlds in which we have P but not Q are metaphysically possible [(from (2a) and (3a)].

(5a) So, the doctrine of physicalism with respect to qualia is incorrect [from (4a) and (1a)].

In other words, the gist of the argument is that zombies are metaphysically possible because we can clearly and coherently conceive of them existing. As such, it follows that the qualitative mental realm doesn't supervene on the micro-physical or explicitly physical realm. So, it follows that the doctrine of physicalism with respect to qualia is false.

As we saw in the previous chapter's discussion of Kripke, the most contentious claim in this argument is the one found in (2a), the idea that conceivability implies metaphysical possibility. Given its contentious character, Chalmers devotes most of his efforts attempting to establish the truth of this claim. In particular, he appeals to a Fregean approach to semantics known as *two-dimensionalism* (2009).

For the two-dimensionalist, words like 'water' – and corresponding concepts like WATER – have two semantic dimensions. On one semantic dimension – the *secondary intension* – 'water' picks out the actual stuff that fills lakes and oceans: H2O. For the semantic movement growing out of Kripke's original account in "Naming and Necessity"

known as *Millianism*[45], this semantic dimension is the only dimension; 'water' just means *H2O*, and refers only to H2O. In terms of possible worlds, the Millian holds that there are no metaphysically possible worlds in which water is something other than H2O. Unlike the Millian, the two-dimensionalist holds that there is another semantic dimension of words like 'water'; this is the *primary intension*. On the primary intension of 'water', the word is synonymous with a non-rigid description like "watery stuff". This intension picks out whatever is epistemically (or – in Chalmers' words – *logically*) possible for water to be (i.e., anything conceivable or not ruled out by logic). For instance, it is uncontroversial to say that water might have been something other than H2O. For the Millian, the way we should think about this possibility is in epistemic terms; once we see that water is actually H2O, we see that this epistemic possibility is not a genuine metaphysical possibility (Soames 2005). For the two-dimensionalist like Chalmers, however, epistemic possibilities such as these are "in the same space of worlds" as other metaphysical possibilities and so they are genuine metaphysical possibilities. To put it in words more in line with Chalmers' account, any logically possible state of affairs corresponds with some world which is genuinely possible. Now, there are two sub-sets of worlds within the space of logically possible worlds – epistemic and metaphysical – but this distinction, for Chalmers, is one that comes from the semantics of primary and secondary intensions. So, there are no epistemically possible worlds that are not also genuinely possible worlds.

[45] From J.S. Mill.

Let us now see how the implications of this framework might bear on the status of the claim that anything ideally conceivable is metaphysically possible (2a). Consider the word 'pain' or its corresponding concept PAIN. Like 'water', the secondary intension of 'pain' picks out whatever it is in the actual world – say *c-fiber firings* – that happens to (in a weak sense) instantiate the property of *being in pain*. The primary intension, which is synonymous with a non-rigid description like "painful things", picks out anything that might conceivably instantiate pains (e.g., ghostly substances). Conversely, in terms of zombies, since we can conceive of c-fiber firings without pains, it seems to follow that this is genuinely possible. (For Chalmers, the same goes for water, as he holds there are worlds where H2O isn't water).

If all of the above is correct, then it follows that all forms of physicalism with respect to qualia are false, since PST (i.e., our commitment to the supervenience of qualitative mental properties on micro-physical properties) is false. As stated before, for the physicalist, any world exactly similar to ours with respect to all the micro-physical facts is exactly like ours with respect to the qualitative mental facts. But, if there are zombie worlds, then there are worlds exactly like ours but lacking qualia and the physicalist is wrong. More specifically for our account, whatever explicitly physical properties "happen" to instantiate qualitative mental properties in our world must instantiate those properties in all genuinely possible worlds (since identity is a necessary relationship). If Chalmers is right, then there are worlds in which we have the firing of c-fiber firings without pains – and worlds in which we have pains without c-fiber firings.

§ 1.2: *Responding to the Two-dimensional Argument*

There are a few ways the physicalist with respect to qualia might respond to Chalmers' argument. One response would be to question some crucial assumptions of the two-dimensional framework, show how they are incorrect, then demonstrate how the anti-physicalist conclusion does not follow. Another response – the one I favor – is to show that even *if* two-dimensionalism is the correct theory of semantics, we are nevertheless unwarranted in deriving such ontological conclusions from it. Here, we shall discuss these two strategies in turn.

§ 1.21: *Rejecting Two-dimensionalism*

One response to Chalmers would be to appeal to a posteriori necessities, like Soames does (as we have seen in the previous chapter). But, if the two-dimensional theory of semantics is correct, then there are no metaphysically necessary truths that we can only come to know a posteriori. Now, we might wonder how this follows, or how a theory concerning semantic and epistemic issues could have such implications. To say, however, that some truth is an a posteriori necessity is to make a metaphysical claim and a claim concerning our *epistemic* relation to that truth. As such, issues concerning the epistemology of modal claims are relevant for an appeal to a posteriori necessities. So, those who appeal to a posteriori necessities to object to Chalmers have a stake in the larger debate concerning the proper semantic theory.

The standard objection to Chalmers' conceivability argument is to claim that our inability to determine a priori that the presence of the set of relevant micro-physical facts necessitates the presence of the set of relevant qualitative mental facts has no bearing on the

truth of PST – since PST is knowable a posteriori, only. That is, all worlds like ours with respect to micro-physical properties are necessarily like ours with respect to qualitative mental properties, but we cannot know that this is the case by a priori reasoning alone. If this is right, then PST may very well be true despite the conceivability of zombies; we do not need to know a priori of p that q follows in order for the conditional relationship between p and q to hold necessarily.

Chalmers objects to this appeal to a posteriori necessities by arguing that knowledge of other worlds can only be obtained a priori, since he holds that a posteriori reasoning can only tell us what is true of the actual world (1996, pg. 137). With the following kind of example, Soames contends that a posteriori reasoning can, indeed, give us knowledge of other worlds (2005, pg. 198). Consider that, if the doctrine of origin essentialism is true, then some facts concerning the origin of a given entity are essential features of that entity. For instance, for the origin essentialist (this is an oversimplification), *having developed from a particular sperm and a particular egg* is an essential property of Saul Kripke; or, in possible worlds talk, there are no possible worlds in which Kripke exists but lacks this property. As Soames points out, we can know a priori of Kripke that having a particular origin is one of his essential properties, while gaining knowledge of facts concerning the particular origin – exactly which sperm and which egg – is an a posteriori affair. By my lights, Soames is correct about this general point against Chalmers and, so, it is plausible that there are at least some necessary truths knowable a posteriori.

Soames is wrong, however, in thinking that this strategy used in the origin essentialism case may be applied to the issue concerning the status of physicalism. Consider that the physicalist with respect to the qualitative mental realm is committed to PST. While it is certainly open for the physicalist to maintain that the property of *being a pain* supervenes on the property of *a particular physical state*, Soames' aforementioned strategy for establishing this would have to go the following, problematic way: *we know a priori of pains that they are essential properties of whatever states that have them, while finding out what kind of states actually have them is an a posteriori matter*. While there might be some plausible arguments for why we can know a priori of a given entity that its origin is essential to it (Forbes 1985, Salmon 2005), it is difficult to see how we can know a priori of pains that they are essential to whatever it is that has them. To put it another way, consider the following argument the physicalist might make along these lines:

(1b) If pains are properties of certain kinds of physical states (leaving it open as to whether or not 'state' is construed globally), they are *essential* properties of those physical states.

(2b) Pains are, in fact, properties of certain kinds of physical states *P*.

(3b) So, pains are essential properties of those kinds of physical states *P*; there are no micro-physical worlds like ours but lacking qualitative mental properties.

In this argument, the physicalist who wants to appeal to a posteriori necessities must first establish the truth of (1b) by a priori means. Yet, (1b) is precisely what Chalmers denies.

One might hope that a posteriori reasoning might settle the score here, but it only comes into play to establish the truth of (2b) – not the truth of (1b).

Someone like Frank Jackson might respond to the above criticism by arguing that, while it is not *prima facie* evident that we can determine a priori that the micro-physical facts necessitate the qualitative mental facts, it is nevertheless the case that an *ideal agent* can determine this to be true a priori (2005). While this might very well be correct, this response does not vindicate the strategy of appealing to a posteriori necessities, since knowing that the state of the micro-physical world necessitates (among other things) the presence of pains is the result of knowing a priori of the micro-physical world that pains are necessitated. That is, we must first know a priori of the micro-physical world in question that pains are essential properties of it. Such an account as this is an a priori route to (3b) and, so, there is not much work for a posteriori reasoning to do.

Another response would be to hold that while this is a problem for the non-reductive physicalist who holds that the relationship between a particular physical state and a particular qualitative mental state is that of (mere) supervenience, this is not a problem for the reductive physicalist who holds that this relationship is that of identity. For the reductive physicalist, it is quite simple to see how one can know a priori *of* (read: *de re*) pains that they are essential to whatever physical state that has them, since this physical state just *is* the qualitative mental state in question. That is, it is obvious that *being a pain* is an essential property of a *particular brain state* if these properties are one and the same.

While this reductive physicalist might be right about these identity claims and the fact that it follows that the relevant essential properties are transitive, it is questionable – like before – that this response vindicates the strategy of appealing to a posteriori necessities. Consider, for example, that we might try – as Soames does – to establish the necessary a posterioricity of the statement "Water is H2O" in the following way: *we can know a priori of water that it is essentially whatever kind of state it is that paradigmatic samples of water have in common; while determining that this kind state is H2O is an a posteriori matter* (2007). Even if this is a plausible account for water, applying this proposed solution to our case is questionable. Firstly, we would need to establish a priori that pains are whatever state it is that paradigmatic instances of pain have in common. Trying to establish this, however, begs the question against Chalmers, since we are trying to establish that pains just *are* the actual states that are responsible for paradigmatic instances of pain, while already holding that pains *must* be whatever these actual states are; but to presume that pains must be physical if they are, as a matter of fact, physical, is to already presume the truth of physicalism.

In response to what has been argued, one might object that it nevertheless seems that a posteriori reasoning plays some part in establishing certain claims about physicalism with respect to qualia. I think this is right but, what is learned a posteriori is not *de re* but *de dicto* (this point will be important later when we discuss the problem of the explanatory gap). For instance, an arguably more plausible reading of what we come to know a posteriori in our pain/c-fiber firings case is that our concepts PAIN and C-FIBER FIRINGS (and their corresponding terms 'pain' and 'c-fiber firings') are co-extensive.

§ 1.22: *From Semantics to Ontology?*

If what I have said in the previous section is correct, then the physicalist has some previously unforeseen difficulties when appealing to a posteriori necessities in order to block Chalmers' inference. All is not lost though, as I am convinced that the solution to this problem is actually quite simple – and requires relatively few ontological or semantic commitments. Instead of responding to Chalmers by appealing to the partly epistemic notion of a posteriori necessities, we might do better to show why his attempt to derive metaphysical/ontological conclusions from semantic/epistemic premises is no more legitimate than any other previous attempts, such as those we saw in the previous chapter.

Soames briefly touches on – but does not flesh out – what I think is the proper response by noting that we can apply the Kripkean strategy of stipulation when trying to make certain metaphysical claims without getting bogged down by semantic issues (2005, chp. 9). Imagine that we want to claim that Aristotle might have been a soldier rather than a philosopher and teacher. The descriptivist might object to this claim by saying that 'Aristotle' just means something like *the teacher of Alexander* and, so, there are no possible worlds in which he exists but lacks this corresponding property. Following Kripke, to respond to this, we needn't have any stake in the semantic debate, since we can simply stipulate that we are talking about the individual referred to by usages of 'Aristotle' in the actual world. That is, as long as we are clear that we are making a *de re* claim – a claim about the individual named 'Aristotle' in the actual world – no semantic issues should bear on our purely metaphysical claim. Likewise, in terms of our case, we can say *of* the properties referred to by 'pains' in the actual world that they are physical whether or not 'pain' might mean something else – or

whether our concept PAIN picks out a different property. Now, Chalmers might respond by saying that if we accept that our concept PAIN might apply to properties other than those to which it refers in the actual world then it follows that pains are not whatever they are in the actual world. But admitting that the *concept* PAIN might be radically disjunctive in this way doesn't mean we must accept that pains, themselves, are disjunctive kinds. To do this would be to read off our ontology from our concepts.

§ 2: THE KNOWLEDGE ARGUMENT

At a first pass, the knowledge argument against physicalism with respect to qualia is the idea that one can know all there is to know about the physical world without knowing all there is to know about the qualitative mental world and, so, it seems to follow that knowledge of, say, the sensation of redness is knowledge of something over and above the physical; hence, physicalism is false. The literature on this argument is vast, as physicalists have devised a large array of responses with the aim of denying the anti-physicalist inference. Instead of surveying this literature in great detail, I shall only briefly sketch the history of this argument and its responses. Instead of exegesis, I shall focus my discussion on the currently most popular physicalist response which might be called the *phenomenal concept strategy* (Stoljar 2005) – henceforth referred to as 'PCS' – or the idea that the purported acquisition of knowledge of the qualitative mental world by acquaintance (e.g., by having a given sensation) is simply the acquisition of concepts (not knowledge of distinctly non-physical facts). While PCS is promising, in light of a certain objection that we shall discuss, it needs some reformulating to work.

§ 2.1: *The Argument in Detail*

Let us examine the knowledge argument in more detail. Discussions of how differences between knowledge claims about the physical world and knowledge claims about the mental world might bear on the status of physicalism might be traced as far back as Descartes, who held that our knowledge of our mental world is incorrigible, while what we take to be knowledge of the external by contrast is fallible. So, he argued, the physical or material world is distinct from the mental world. More modern discussions can be found in Herbert Feigl's discussion (1958) of the possibility of aliens who know everything about human physiology but nothing of human experience, and Thomas Nagel's discussion (1974) of the fact that we can know of bats what Feigl's aliens know of us but lack knowledge of *what it is like* to be a bat.

It is not until Jackson's discussion (1986) of this problem that we get an explicit argument against physicalism, so 'the knowledge argument' is generally taken to refer to Jackson's argument specifically, despite sharing insights with previous discussions. Jackson's initial formulation of the argument is the thought experiment contained in the following passage:

> Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor. She specializes in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red', 'blue', and so on. She discovers, for example, just which wavelength combinations from the sky stimulate the retina, and exactly how this produces *via* the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue'.… What will happen when Mary is released from her black and

> white room or is given a color television monitor? Will she *learn* anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then is it inescapable that her previous knowledge was incomplete. But she had *all* the physical information. *Ergo* there is more to have than that, and Physicalism is false (pg. 130).

This argument has a lot of intuitive pull, as it certainly seems that Mary not only learns something when she leaves her monochrome room but learns a new fact. That is, it certainly seems that knowledge of – to use Nagel's phrase – what it is like to experience, say, the sensation of redness is knowledge of something that is a fact outside the set of all physical facts. If this is right, it certainly seems that qualia must be non-physical properties.

There are several ways to formulate the argument contained within this passage more explicitly; but we may formulate it the following way to ensure that an ontological claim, rather than an epistemic claim follows as the conclusion:

(1c) Prior to leaving her room, Mary knows all the physical facts about vision.

(2c) After leaving her room, Mary acquires knowledge about a new fact (e.g. what it is like to experience redness).

(3c) So, there are non-physical facts about vision; that is, physicalism is false.

Formulated this way, the argument makes it clear that Mary purportedly learns something about a new fact. Since it is held that she previously knows all the physical facts, it follows that she learns something about a non-physical fact.

Some have argued that this argument is a non-starter just like other arguments purporting to derive ontological conclusions from epistemic conclusions. But, as Robert

Stalnaker notes[46], unlike other epistemic claims, a claim that one knows that P implies the truth of P; so, we can, indeed, derive substantive conclusions in the way that Jackson does in this argument. So, the knowledge argument is one that we physicalists should take seriously.

§ 2.2: *The Phenomenal Concept Strategy*

As stated before, the literature on the knowledge argument is vast. So, instead of spending the tens (or hundreds) of pages needed to give all of these responses fair representation, I shall focus on what is known as *the Phenomenal Concept Strategy* (PCS).

At first, those appealing to the strategy in question attempt to give a reading of the Mary case that is consistent with the doctrine of physicalism with respect to qualia. The strategy is as follows. Instead of acquiring new knowledge such as in (2c), it is argued that Mary acquires a new concept – namely, a phenomenal concept. So, it is held that (2c) is false and, so, (3c) doesn't follow – and physicalism remains intact. Indeed, it should be noted further that with this strategy, (1c) is false, since there are some facts concerning concepts about vision that Mary does not know.

As Derek Ball (2009) and Michael Tye (2009) note, the appeal to phenomenal concepts relies on a fine-grained, Fregean conception of concepts, where concepts are individuated by something other than their referents. For example, it might be held that prior to leaving her room, Mary has the concept PCA (pyramidal cell activity in the primary visual cortex), while after leaving the room and seeing a red object she acquires the concept REDp (where 'p' designates that this concept is phenomenal). From this, it is argued that the

[46] From the John Locke Lectures (2007), lecture 2.

acquisition of a new concept is not the acquisition of knowledge concerning a new fact, as facts are individuated in a more coarse-grained fashion.

§ 2.21: *Problems with PCS*

While PCS has intuitive pull, it has some problems. As philosophers of language have increasingly come to accept the doctrine of semantic externalism – the idea that semantic content is determined by representational content – with respect to (minimally) singular terms such as the proper name 'Aristotle' and general terms such as 'Water', philosophers of mind have generally followed suit and have come to accept that mental content is also exhausted by representational content (Fodor 2008, Edwards 2010). For example, as Hilary Putnam and Tyler Burge have forcefully argued, concepts like ELM are not individuated by the properties that come to one's mind through introspection; rather, these concepts are individuated by external factors and possession of these concepts is deferential.

Specifically, with respect to the knowledge argument and the problem of qualia, Michael Tye and Derek Ball have argued that we should think that if there are phenomenal concepts, like other concepts, they should be individuated like other concepts are. But, since it seems that phenomenal concepts must be individuated in an intensional, Fregean fashion, it follows that there are no phenomenal concepts. As such, those appealing to PCS can no longer hold that Mary acquires a new, phenomenal concept, as there are no such things.

Ball makes the further claim that the physicalist needn't be worried about the non-existence of phenomenal concepts, since – as he argues (I think, rightly) – the knowledge argument itself relies on there being phenomenal concepts. To see this, consider that the

proponent of the knowledge argument must be committed to the following claim: Mary cannot know what it is like to see red without having the concept REDp. Having REDp is a necessary condition for know what it is like to see red. So, if there is no REDp, then Mary does not acquire knowledge about what it is like to see red.

§ 2.22: *Problems With Arguments Against PCS*

Let us agree with Ball and Tye and grant that conceptual content just is representational content. The argument that the truth of this claim implies that there are no phenomenal concepts might be construed as follows: if there are phenomenal concepts, they must be Fregean; no concepts are Fregean; so, there are no phenomenal concepts. The support for the claim that phenomenal concepts must be Fregean, it seems, is simply an appeal to what those appealing to PCS take them to be. For instance, Ball rightly argues that those appealing to PCS adhere to what he calls the phenomenal concept criterion – or PCC – which is the idea that a token of a given concept is a phenomenal concept only if that token is instantiated by one who has had the experience of the relevant qualitative state (pg. 938). For example, those adhering to PCC would hold that one only has the phenomenal concept REDp if one has previously had the sensation of redness.

Even if we grant that those who appeal to PCS also appeal to PCC, it doesn't follow that those appealing to PCS must appeal to PCC (at least as it has been construed thus far). If we are externalists about conceptual content, we hold that we may be wrong about what a given concept requires. For example, in Burge's ARTHRITIS case, it is clear that one might have an inaccurate conception of what arthritis is, while nevertheless being a competent user

of the concept ARTHRITIS (1979). Likewise, it may be the case that those who adhere to PCC are just wrong about what is required for a concept to be a phenomenal concept. Ironically, for Tye and Ball to deny this is to accept, in some form or other, a Fregean view of concepts, as they allow that the intension (as determined by those appealing to PCS) of concepts like REDp determines the extension.

If all of the above is correct, then it is still an open question whether or not there are phenomenal concepts. Let us grant that one may have the concept REDp without having experienced the sensation of redness. While Tye and Ball are correct to argue that concepts, in general, are not Fregean, it certainly seems that there is something to the claim that there are phenomenal concepts. How might we do justice to this intuition without betraying both physicalism and externalism? Let us start by weakening PCC with the following reformulation PCC`: a concept C is a phenomenal concept if and only C refers to some set of qualitative mental properties P, where the reference relation between C and P is either actual or possible. So, imagine that John blows a dog whistle to get the attention of his pet dog. John has never heard a dog whistle and it is unlikely that his imagining what it is like to hear a dog whistle accurately captures the content of the experience. Despite this, John is a competent user of the concept DOG-WHISTLE-SOUND (where 'sound' is construed subjectively) because the corresponding linguistic phrase, say, "the sound a dog whistle makes" refers to an actual token of this qualitative mental property that happens to be instantiated in the mind of his dog. In a more counterfactual mood, imagine a scenario where Jill is trying to imagine what it is like to feel a pain worse than any creature in the

actual world has ever felt – or can ever feel. It seems plausible that, in this case, though there is no actual referent of the concept PAIN WITH X INTENSITY, it seems that Jill nevertheless has this concept since it picks out an at least epistemically possible qualitative mental state, if not a metaphysically possible one.

If there are phenomenal concepts in the weak sense, as sketched above, then it seems that we might (again) have a problem when responding to the anti-physicalist. Recall that Ball argues that the physicalist needn't worry if there are no phenomenal concepts, since the knowledge argument, itself, relies on there being such concepts. If what I have argued is correct, however, the non-physicalist can agree that phenomenal concepts are non-Fregean and referential; they would just hold that these concepts are individuated by their non-physical referents.

Holding that phenomenal concepts are referential doesn't do the non-physicalist much good, however. To make this point clear, consider that those appealing to the knowledge argument in order to establish that qualia are non-physical would have to establish the following: prior to leaving the room, Mary has the concept PCA, but acquires the **new** concept REDp upon leaving, since the sensation of redness is non-physical and, hence, numerically distinct from pyramidal cell activity. The problem here is that the only reason that we would accept that Mary acquires a new concept is if we already accept that the kinds of properties picked out by the concepts are numerically distinct.

§ 2.23: *Another Try With The Phenomenal Concept Strategy*

With PCC` in hand, an alternate reading of Mary's situation is available to the physicalist: upon leaving the room, Mary acquires no new concept; rather, a token of the same concept type – REDp – is instantiated. That is, the physicalist can say that Mary already had the concept REDp before leaving the room though, perhaps, she did not fully understand it; but lack of full understanding, as Tye notes, does not bar one from being a competent user of that concept, since concepts are deferential.

Construed in the above way, Mary's situation is not unlike someone in the following situation. Imagine that Barry came to know everything about the property being a sample of H2O at a micro-physical level prior to coming in contact with the seemingly emergent property being a sample of water. Since water just is H2O, all the facts concerning H2O, concern water, as well. So, if Barry knows something of H2O, he knows the same thing of water. Now, after coming in contact with water at a macro-physical level, it certainly seems that Barry acquires knowledge of a new fact, and this is right in some sense; but this is knowledge *de dicto*. That is, Barry learns a new way of describing the very same phenomenon. Now, one might reply that there is a difference between linguistic items and mentalistic items and, so, descriptions are not part of a mentalistic ontology. But, if we accept the language of thought hypothesis (Fodor 1975), then what goes for public language might plausibly be maintained to go for the language of thought. So, in Mary's case we can say that there are two ways of describing the same phenomenon in the language of thought: as PCA and as REDp, where knowledge that these two concepts are actually tokens of the same concept is acquired a posteriori; and this knowledge is *de dict*o.

§ 3: IMPLICATIONS FOR THE EXPLANATORY GAP

Even if we accept that, at the end of day, qualia are physical properties, one might object that we cannot understand how this might be the case. That is, as McGinn argues, it seems that the chasm between our understanding of the physical and our understanding of the qualitatively mental – the explanatory gap – is unbridgeable, in principle. If he is right, it seems that consciousness will forever be a mystery. This claim, if true, would be unsettling; further, it would mean that our efforts to try to understand (at least in some sense of 'understand') the conscious mind in physical terms are in vain.

McGinn's argument for why the explanatory gap is unbridgeable relies on his claim that we just aren't wired to have the right kind of concepts required to understand the phenomenon. Echoing this sentiment (though he thinks the gap is unbridgeable because the problem is ontological), Chalmers argues that our conception of the physical world is in terms of the structure and dynamics of physical entities, while our conception of the qualitative mental world is in distinctly phenomenological terms; so, we cannot deduce the phenomenological structure of the mental world from the structure of the physical world (2003). In other words, while we might very well have what we have called a fundamental explanation of why qualia arise in our brains (i.e., an account of that in virtue of which qualia arise in our brains), it seems that we might nevertheless lack what we have called a pragmatic explanation for why this is (i.e. an explanation that is described in a way that is intelligible to us).

In reply to McGinn and Chalmers, we might say that, if we accept that concepts are individuated referentially, it follows that we aren't lacking any important concepts needed to

understand how qualia arise in our brains. Instead, if what we have argued thus far is correct, what we lack is knowledge concerning how to translate one description in the language of thought into another – a description of qualia in a physicalistic vocabulary into a description in a folk (in this case, phenomenological) vocabulary. Doing this seems difficult, but translations like this actually occur quite frequently. For example, if a physicist is trying to explain what a superposition is, or what it means to say we live in eleven dimensions, she uses certain analogies, appealing to phenomena of which we have a good understanding. The reason for doing this is because we have two different ways of understanding the world: a fundamental way (e.g. in terms of formulas) and pragmatic way (e.g. in terms of models).

Now, one could object to our appeal to the physicist's situation by saying that qualia just aren't like anything else in the world, so we cannot construct the same kind of analogies as we do to explain physical phenomena. This seems incorrect to me. Indeed, that there has already been some progress in attempting to explain what qualia are like. For instance, Douglas Hofstadter has recently likened conscious phenomena to strange loops, the seemingly emergent, recursive phenomenon that happens when certain kinds of devices (e.g. video cameras) turn inward and represent themselves (2006). Now, this account might not be correct at a fundamental level, but it does suggest that there might be a way of understanding qualia in an intelligible way, after all. Indeed, I speculate that, as we come to understand what laws govern qualia, we will find better ways of understanding these properties in a pragmatic way.

CONCLUSION

This chapter was a response to anti-physicalist arguments, in general.

In Section 1, we looked at Chalmers' conceivability argument against the doctrine of physicalism with respect to qualia. Chalmers maintains that conceivability implies genuine possibility because of his two-dimensionalist semantic framework. Soames' response to this relies on a posteriori necessities. However plausible this might be for essential origin properties, as I argued, it does not work for us, since it would be begging the question to presume that pains are whatever actually constitute them. From here, I argued that the best way to respond to the two-dimensionalist strategy is to elucidate how the semantics of a term is supposed to determine what is metaphysically possible, as Chalmers would have it. On close inspection the suggestion that semantics determines what is metaphysically the case just can't be plausibly maintained – just like it couldn't be plausibly maintained against Smart, half a century before.

In Section 2, we looked at the knowledge argument against physicalism. The popular strategy of appealing to phenomenal concepts, as we have seen, cannot so easily be maintained in light of the externalist consensus on concepts in general. We can, however, as I argued, modify this strategy in such a way that we can maintain that phenomenal concepts, like other concepts, are deferential.

In Section 3, we looked at the explanatory gap, or the problem concerning how it is that the stuff that constitutes the brain could possibly be the same stuff that constitutes the mind. Drawing from the distinction drawn in the second chapter – that between fundamental and pragmatic explanations – I made the case that bridging the gap can be done

by translating our seemingly incommensurate vocabularies in the same way that scientists do in other domains.

WORKS CITED

Albert, D. Z. (2000). Time and Chance, Harvard University Press.

Ball, D. (2009). "There are no phenomenal concepts." Mind **118**(472): 935-962.

Bechtel, W. P. and J. Mundale (1999). "Multiple realizability revisited: Linking cognitive and neural states." Philosophy of Science **66**(2): 175-207.

Bennett, K. (2003). "Why the exclusion problem seems intractable and how, just maybe, to tract it." Noûs **37**(3): 471-497.

Bennett, K. (2008). Exclusion again. Being Reduced: New Essays on Reduction, Explanation, and Causation. J. Hohwy and J. Kallestrup, Oxford University Press.

Bickle, J. (1997). Psychoneural Reductionism: The New Wave, MIT Press.

Bickle, J. (2005). "Precis of _Philosophy and Neuroscience: A Ruthlessly Reductive Account." Phenomenology and the Cognitive Sciences **4**(3): 231-238.

Block, N. (forthcoming) "Functional reduction" in D. Sosa, T. Horgan and M. Sabatés (eds) Supervenience in Mind: A Festschrift for Jaegwon Kim. Cambridge, Mass.: MIT Press.

Block, N. (1990). "Inverted earth." Philosophical Perspectives **4**: 53-79.

Block, N. (1996). What is functionalism? [Book Chapter]. D. M. Borchert, MacMillan.

Block, N. and J. A. Fodor (1972). "What psychological states are not." Philosophical Review **81**(April): 159-181.

Block, N. and R. Stalnaker (1999). "Conceptual analysis, dualism, and the explanatory gap." Philosophical Review **108**(1): 1-46.

Burge, T. (1979). "Individualism and the mental." Midwest Studies in Philosophy **4**(1): 73-122.

Campbell, N. (2000). "Physicalism, qualia inversion, and affective states." Synthese **124**(2): 239-256.

Carnap, R. (1932). "Psychologie in physikalifcher sprache." Erkenntnis **3**(1).

Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory, Oxford University Press.

Chalmers, D. J. (2003). Consciousness and its place in nature. Blackwell Guide to the Philosophy of Mind. S. P. Stich and T. A. Warfield, Blackwell.

Chalmers, D. J. (2009). The Two-Dimensional Argument Against Materialism. Oxford Handbook to the Philosophy of Mind. B. P. McLaughlin and S. Walter, Oxford University Press.

Churchland, P. M. (1985). "Reduction, qualia and the direct introspection of brain states." Journal of Philosophy **82**(January): 8-28.

Crick, F. and C. Koch (1990). "Toward a neurobiological theory of consciousness." Seminars in the Neurosciences **2**: 263-275.

Davidson, D. (1970). Mental events. Experience and Theory. L. Foster and J. W. Swanson, Humanities Press**:** 79-101.

Davidson, D. (2001). Essays on Actions and Events: Philosophical Essays Volume 1, Clarendon Press.

Dennett, D. C. (1988). Quining qualia. [Book Chapter]. A. J. Marcel and E. Bisiach, Oxford University Press.

Dennett, D. C. (1991). Consciousness Explained, Penguin.

Edwards, K. (2010). "Concept referentialism and the role of empty concepts." Mind and Language **25**(1): 89-118.

Evans, J. and K. Frankish (2008). In Two Minds: Dual Processes and Beyond, Oxford University Press.

Feigl, H. (1958). "The 'mental' and the 'physical'." Minnesota Studies in the Philosophy of Science **2**: 370-497.

Flanagan, O. J. (1984). The Science of the Mind, MIT Press.

Flanagan, O. J. and T. W. Polger (1995). "Zombies and the function of consciousness." Journal of Consciousness Studies **2**(4): 313-321.

Fodor, J. A. (1974). "Special Sciences (Or: The Disunity of Science as a Working Hypothesis)." Synthese **28**(2): 97-115.

Fodor, J. A. (1975). The Language of Thought, Harvard University Press.

Fodor, J. A. (1986). The modularity of mind. Meaning and Cognitive Structure. Z. W. Pylyshyn, Ablex.

Fodor, J. A. (1991). "You can fool some of the people all of the time, everything else being equal: Hedged laws and psychological explanation." Mind **100**(397): 19-34.

Fodor, J. A. (1997). "Special sciences: Still autonomous after all these years." Philosophical Perspectives **11**: 149-163.

Fodor, J. A. (2008). Lot 2: The Language of Thought Revisited, Oxford University Press.

Forbes, G. (1985). The Metaphysics of Modality, Clarendon Press.

Hardcastle, V. G. (1997). "When a Pain is Not." The Journal of Philosophy **94**(8): 381-409.

Heil, J. (2003). From an Ontological Point of View, Oxford University Press.

Hill, C. S. (1991). Sensations: A Defense of Type Materialism, Cambridge University Press.

Hill, C. S. (1997). "Imaginability, conceivability, possibility, and the mind-body problem." Philosophical Studies **87**(1): 61-85.

Hill, C. S. (2009). Consciousness, Cambridge University Press.

Hofstadter, D. R. (2006). What is it like to be a strange loop? Self-Representational Approaches to Consciousness. U. Kriegel and K. Williford, MIT Press.

Horgan, T. E. (2001). "Causal compatibilism and the exclusion problem." Theoria **16**(40): 95-116.

Horowitz, A. (1999). "Is there a problem in physicalist epiphenomenalism?" Philosophy and Phenomenological Research **59**(2): 421-434.

Humphrey, N. (1974). "Vision in a monkey without striate cortex: A case study." Perception **3**(3): 241-255.

Humphrey, N. (1992). A History of the Mind: Evolution and the Birth of Consciousness, Simon and Schuster.

Humphrey, N. (2006). Seeing Red: A Study in Consciousness, Belknap Press.

Huxley, T. (1874). "On the hypothesis that animals are automata, and its history." Fortnightly Review **95**: 555-580.

Jackson, F. (1982). "Epiphenomenal qualia." Philosophical Quarterly **32**(April): 127-136.

Jackson, F. (1986). "What Mary didn't know." Journal of Philosophy **83**(May): 291-295.

Jackson, F. (2005). The Case for a Priori Physicalism. Philosophy-Science -Scientific Philosophy, Main Lectures and Colloquia of Gap 5, Fifth International Congress of the Society for Analytical Philosophy. C. Nimtz and A. Beckermann, Mentis.

Kim, J. (1998). Mind in a physical world : an essay on the mind-body problem and mental causation. Cambridge, Mass., MIT Press.

Kim, J. (2003). Philosophy of mind and psychology. Donald Davidson, Cambridge: Cambridge University Press.

Kim, J. (2005). Physicalism, or Something Near Enough, Princeton University Press.

Kripke, S. A. (1980). Naming and Necessity, Harvard University Press.

Le Bihan, D., et al. (1993). "Activation of human primary visual cortex during visual recall: a magnetic resonance imaging study." Proceedings of the National Academy of Sciences **90**(24): 11802-11805.

Levine, J. (1983). "Materialism and qualia: The explanatory gap." Pacific Philosophical Quarterly **64**(October): 354-361.

Loewer, B. (2002). "Comments on Jaegwon Kim's mind and the physical world." Philosophy and Phenomenological Research **65**(3): 655–662.

Loewer, B. (2008). "8. Why There Is Anything Except Physics." Being Reduced **1**(9): 149-164.

Loewer, B. (2009). "Why is there anything except physics?" Synthese **170**(2): 217 - 233.

McGinn, C. (1989). "Can we solve the mind-body problem?" Mind **98**(July): 349-366.

McLaughlin, B. P. (1989). "Type epiphenomenalism, type dualism, and the causal priority of the physical." Philosophical Perspectives **3**: 109-135.

McLaughlin, B. P. (1992). On Davidson's response to the charge of epiphenomenalism. Mental Causation. J. Heil and A. R. Mele, Oxford University Press.

Nagel, T. (1974). "What is it like to be a bat?" Philosophical Review **83**(October): 435-450.

Ney, A. (2007). "Can an appeal to constitution solve the exclusion problem?" Pacific Philosophical Quarterly **88**(4): 486–506.

Place, U. T. (1956). "Is consciousness a brain process?" British Journal of Psychology **47**(1): 44-50.

Polger, T. W. (2008). "Two Confusions Concerning Multiple Realization." Philosophy of Science **75**(5): 537-547.

Polger, T. W. (2011). "Are sensations still brain processes?" Philosophical Psychology **24**(1): 1-21.

Polger, T. W. and O. J. Flanagan (1996). "Explaining the evolution of consciousness: The other hard problem."

Popper, K. R. and J. C. Eccles (1977). The Self and Its Brain: An Argument for Interactionism, Springer.

Pryor, J. (2000). "The skeptic and the dogmatist." Noûs **34**(4): 517–549.

Putnam, H. (1967). Psychological predicates. Art, Mind, and Religion. W. H. Capitan and D. D. Merrill, University of Pittsburgh Press.

Putnam, H. (1975). "The meaning of 'meaning'." Minnesota Studies in the Philosophy of Science 7: 131-193.

Ramachandran, V. S. and W. Hirstein (1998). "Three laws of qualia: What neurology tells us about the biological functions of consciousness." Journal of Consciousness Studies 4(4-5): 429-457.

Ryle, G. (1949). The Concept of Mind, Hutchinson and Co.

Salmon, N. U. (2005). Reference and Essence, Prometheus Books.

Schlick, M. (1935). De la relation entre les notions psychologiques et les notions physiques. Die Wiener Zeit, Springer Vienna. **6:** 575-609.

Shiffrin, R. M. and W. Schneider (1984). Automatic and controlled processing revisited.

Shoemaker, S. (1982). "The inverted spectrum." Journal of Philosophy **79**(July): 357-381.

Smart, J. J. C. (1959). "Sensations and brain processes." Philosophical Review **68**(April): 141-156.

Smart, J. J. C. (1961). "Colours." Philosophy **36**(April-July): 128-142.

Smart, J. J. C. (1967). "Comments on the Papers."

Smart, J. J. C. (2006). "Metaphysical illusions." Australasian Journal of Philosophy **84**(2): 167 – 175.

Soames, S. (2005). Reference and Description: The Case Against Two-Dimensionalism, Princeton: Princeton University Press.

Soames, S. (2007). "What Are Natural Kinds?" Philosophical Topics **35**(1-2): 329-342.

Stoljar, D. (2005). "Physicalism and phenomenal concepts." Mind and Language **20**(2): 296-302.

Tye, M. (2009). Consciousness Revisited: Materialism Without Phenomenal Concepts, Mit Press.

Weiskrantz, L. (1986). Blindsight: A Case Study and Implications, Oxford University Press.