A Physicalist Solution to the Explanatory Gap

by

Yanssel Garcia

Submitted in Partial Fulfillment of the

Requirements for the Degree

Doctor of Philosophy

Supervised by Professor Earl Conee

Department of Philosophy
Arts, Sciences, and Engineering
School of Arts and Sciences

University of Rochester
Rochester, New York

2021

Table of Contents

Biographical Sketch iii

Abstract iv

Contributors and Funding Sources v

Chapter 1 1

Chapter 2 76

Chapter 3 125

Chapter 4 173

Bibliography 223

Biographical Sketch

Yanssel Garcia was born in Miami, Florida, USA. He attended Florida International University in 2010, and he graduated with a Bachelor of Arts degree in Philosophy in 2014. In 2015, he began his doctoral studies in Philosophy at the University of Rochester. He was awarded the Provost Fellowship that same year. He received his Masters degree in 2020 at the University of Rochester, and he has pursued his research in Philosophy of Mind under the direction of Earl Conee.

Abstract

As substance dualism fell out of favor, philosophers became increasingly interested in making sense of mind in purely physicalist terms. Along the way, the physicalist project has hit a few snags. Perhaps the most popular challenge was presented by Frank Jackson's Mary's Room thought experiment, wherein Mary, a brilliant color scientist, comes to know all of the physical facts about color whilst confined to a black-and-white room. Once released, Mary is presented with a ripe tomato. The intuition is that Mary, upon seeing a colored object for the first time, has learned something new, but what she has learned apparently cannot be accounted for by physicalism, thereby leaving an explanatory gap between mind and matter. There are those, like Joseph Levine, who believe the explanatory gap to be a necessary consequence of any physicalist theory of mind. I disagree, and in this dissertation, I aim to show that at least one physicalist theory of mind can close the gap. However, it requires embracing a theory that physicalists are hesitant to embrace: panpsychism.

Contributors and Funding Sources

This work was supported by a dissertation committee consisting of Professor Earl Conee and Professor Paul Audi of the Department of Philosophy and Professor John Heil of the Department of Philosophy of the University of Washington at St. Louis. Graduate study was supported by the Provost Fellowship.

Chapter One

Section A: The Explanatory Gap

Where do minds fit into our picture of the world? Substance dualists believe that the mind and the body are two fundamentally distinct substances that share an intimate relationship. The mind, on this view, is an immaterial soul, and this soul is capable of interacting with its given body. Substance dualism has largely fallen out of favor for a variety of reasons. One such reason is that the interaction between material and immaterial substances is deeply mysterious. Of perhaps greater import, though, has been the remarkable success of the sciences in making sense of once deeply mysterious phenomena in purely physical terms. Given this success, many believe that it is only a matter of time before we finally make sense of the mysterious relationship between mind and body. Indeed, J. J. C. Smart cites this very belief as the primary motivation for the identity theory, which states that mental phenomena, which he calls 'sensations', are identical with physical states of the brain, or 'brain processes'. He characterizes the identification of the mental with the neural as a necessary application "of Occam's razor," for the alternative leaves mentality as the one thing in this world that is not subject to physical explanation (Smart, 1959: 142).

Smart argues that the relationship between mental processes and physical processes is one of strict identity (1959: 144). It's worth taking a second to elucidate what this means. Consider water as we were familiar with it four hundred years ago. A complete understanding of water included knowing that it was the clear liquidy stuff that fell from the skies as rain and

was present in the oceans. At some point, we came to discover what water is essentially: we discovered that water is H_2O. In other words, water is strictly identical with H_2O. Note that this is not to say that 'water' and 'H_2O' mean the same thing—the concept of water and that of H_2O are clearly different: When we think of H_2O, a particular molecular structure comes to mind, whereas the thought of water brings along all of those mental images of rain, rivers, and oceans. Nonetheless, the term 'water' and the term 'H_2O' have the same referent—they pick out the same thing in the world. Similarly, the identity theorist admits that minds and brains are conceptually different—one concerning thoughts and feelings and the other a peculiar material object—but they are in actuality the same thing; the term 'brain' and the term 'mind' have the same referent.[1]

Smart provides a number of points in defense of the identity theory. The heart of his defense is the success of the sciences in providing physical explanations for all other phenomena, meaning the mind and body must be subject to the same type of explanation (Smart, 1959: 142). We know that there appear to be highly intimate correlations between mental events and neural events, and Smart believes the best explanation is that these aren't mere correlations, but that the two are the same event. In order to believe that the mental is distinct from the physical, he thinks we would require psychophysical laws that explain the interaction between them. These psychophysical laws, however, would have to "relate simple constituents to configurations consisting of perhaps billions of neurons," and "such ultimate laws would be" highly disjunctive and "like nothing so far known in science," so Smart

[1] This is only roughly true. Rather, any given mental predicate refers to the same thing in the world as some physical (neural) predicate

dismisses the possibility, though the only defense of this move is that such laws "have a queer 'smell' to them" (1959: 143). Additionally, Smart believes that allowing the mental as a separate, epiphenomenal entity would be similar to the theory that the world was created in 4004 B.C. precisely as we find it. Such a theory is completely unfalsifiable and involves "too many brute and inexplicable facts" (1959: 155).

I find Smart's various defenses of the identity theory faulty for a number of reasons. Briefly, the fact that the mental should find its place in the realm of the physical does not entail that the identity theory is the only available option. Furthermore, the identity theory suffers from a flaw similar to the one Smart attributes to the disjunctive, complex nature of psychophysical laws, since the identities it draws relate highly complex material arrangements with simple sensations, which is at least equally mysterious. These identities are brute and rampant, since there must be a brute identity for each and every possible mental state with some physical state. As such, the complex phenomenal state in which I currently find myself is identical to some brain state, and this identity is just a fact of the universe.

These considerations against the identity theory may be too quick, and the identity theorist is not without defense. We'll return to a closer examination of the view later on in this chapter. For now, what matters is that we have a rough characterization of a physicalist attempt at explaining the relation between the mind and body. Joseph Levine identifies a far more worrisome problem with the identity theory in his paper "Materialism and Qualia: The Explanatory Gap," though the problem he identifies is significantly more far-reaching than the identity theory. It is this problem with which I wish to primarily concern myself. The purpose

of this dissertation is to provide a theory that adequately solves Levine's problem. First, however, we must come to grips with what the problem is.

According to Levine, there exists a gap in explanation between physical and mental facts. In his words: "Psychophysical identity statements leave a significant explanatory gap, and, as a corollary, [...] we don't have any way of determining exactly which psychophysical identity statements are true" (Levine, 1983: 354). In other words, it seems that if the identity theory is true, it remains opaque why a mental state is identical with the physical state it's identical with. The problem is an epistemic version of a similar metaphysical problem raised by Saul Kripke in his *Naming and Necessity*. According to Levine, the problem raised by the explanatory gap is deeply troubling. To motivate exactly what's so troubling about the explanatory gap, we should turn to Kripke's original version.

Kripke's original problem concerns the apparent contingency of certain identity statements. Consider the following two sentences:

(1) Pain is the firing of C-fibers.

(2) Heat is the motion of molecules.

Statements (1) and (2) strike us as contingent. Consider (2). It seems, at least at first glance, that heat could have failed to be the motion of molecules. However, Kripke explains, the apparent contingency of (2) can be explained away. We normally associate heat with the familiar phenomenal experience of warmth. But once we clarify that by 'heat' we simply mean 'that which is responsible for x, y, and z', the apparent contingency dissipates. Heat is what's responsible for water boiling, cheese melting, and, of course, the familiar sensation of warmth.

It turns out, as science tells us, that the stuff responsible for these various phenomena is the motion of molecules. Once we understand this, we come to realize that not only is heat the motion of molecules, there's nothing contingent about it. In any world (with the same physical laws) it would remain the case that heat—the motion of molecules—is precisely what's responsible for water boiling and cheese melting.

Now consider (1). Employing the same method used above fails. A world with pain but with no C-fibers is simply a world in which pain exists without C-fibers. Whereas it's impossible upon an adequate understanding of the terms to imagine a world with heat but without molecular motion, we can easily imagine a world with pain but without C-fibers. When it comes to pain, "we cannot make the distinction here, as we can with heat, between the way it appears to us and the phenomenon itself" (Levine, 1983: 355). We solve the problem with (2) by separating the sensation of warmth from what we mean by 'heat'; we utilize an objective understanding of heat that is separate from the phenomenal experience. We cannot do the same thing with pain.[2] Pain just is the phenomenal experience itself.

Kripke believes that these considerations have dire consequences for the identity theory. The terms 'pain' and 'C-fibers' are rigid designators. Rigid designators pick out the same thing in all worlds. This is as opposed to descriptions, which can pick out different entities in different worlds. For instance, the rigid designator 'Barack Obama' picks the same individual across worlds, whereas the description 'the 44th president of the United States'

[2] This gap in explanation exists specifically between qualitative mental states, such as pain, and physical states. No such problem seems to arise between functionally analyzable states such as beliefs and desires. Throughout the dissertation, when I use terms such as 'mind' and 'mental', I am specifically concerned with these qualitative states unless otherwise specified.

would presumably pick out different individuals in different worlds. Now, identity statements between rigid designators are necessarily true if they're true at all (Kripke, 1980: 108). Consider Hesperus and Phosphorus. 'Hesperus' and 'Phosphorus' are rigid designators. 'Hesperus' picks out the planet Venus, as does 'Phosphorus', and in this world 'Hesperus is Phosphorus' is true. Thus, since the statement is true, then it is true in any world that contains Venus. Similarly, 'pain' and 'C-fibers' are rigid designators. If the identity theory is true, then the statement 'pain is C-fibers firing' is true. If the statement is true, then it is necessarily true. However, we're generally convinced that pain is not *necessarily* C-fibers firing, given the ease with which we can imagine a world with pain without C-fibers. Therefore, pain is not C-fibers firing, and so the identity theory is false. This is the reasoning in very crude form; we'll consider a more careful version of this argument later on.

Levine believes that the metaphysical conclusion is too strong. The identity theorist can say that our intuitions are faulty and that the apparent contingency is merely that: apparent. In actuality, pain really could be only C-fibers firing, and so the statement 'pain is C-fibers firing' is necessarily true. This is certainly a possibility. Levine opts for a weaker, but still deeply troubling, version of Kripke's problem: it may turn out that (1) is necessarily true, but no explanation could possibly make sense of it. This counts as a serious point against the identity theorist, and, as we'll see shortly, physicalist theories of mind more generally.

That the identity theory not only doesn't provide an explanation but actually cannot provide one is problematic. This renders the felt contingency of statements like (1) necessary: there will never be anything that makes (1) feel necessarily true. These brute identities between

the physical and the mental appear unacceptably mysterious, and the identity theory provides no promise of ever making sense of them. This lack of explanatory power is troublesome, and a physicalist theory that can avoid this problem would be preferable. However, as Levine makes clear, the identity theory is not the only victim of the explanatory gap: functionalism is affected as well.

Functionalism attempts to avoid the issue by abstracting mental phenomena away from their physical realizers. Thus, pain won't be identified with the firing of C-fibers, but rather with some functional state that's simply realized by C-fibers but could equally be realized by any other number of physical arrangements. This means that phenomenal states are not identified with physical states, but rather identified with functional states. Consider the following statement.

(3) To be in pain is to be in State F.

This kind of identification fares no better, claims Levine, for "it seems imaginable that in some possible world (perhaps even in the actual world) (3) is false" (1983: 356). Block, Levine mentions, provides a persuasive example where the realizer of some mental state turns out to be the entire nation of China fulfilling the same functional role that our brains fulfill when we're in the same mental state (1978: 279). Briefly, it would be possible to have the citizens of China realize State F. We could give each individual a walkie-talkie and have them send signals to one another in the exact same way that the human brain does when it is in, say, pain. It follows from (3) that the nation of China, given its current functional state, is in pain. But it *seems* obvious that the nation of China is not experiencing pain or really anything at all, making this

a clear counterexample to statements such as (3).[3] Even supposing it were sensible to grant mentality to the nation of China thus arranged, we still lack an explanation for the relation between the sensation of pain and the functional state F. We could still make sense of the functional state itself playing the relevant causal roles without making sense of whether the associated feeling is a pain or a tickle. Once again, it appears that no amount of physical information could dispel the felt contingency of phenomenal/physical identifications.

Broadly speaking, there are two major problems that the explanatory gap poses. The first is the deep contingency appealed to above—the complete lack of a satisfying explanation. Let's expand. Reconsider statement (2), which claims that heat is molecular motion. Levine asks: what is it that's explanatory about statement (2)? The explanatory force of (2), says Levine, is captured by statement (2') below (1983: 357):

> (2') The phenomenon we experience through the sensations of warmth and cold, which is responsible for the expansion and contraction of mercury in thermometers, which causes some gases to rise and others to sink, etc., is the motion of molecules.

According to Levine, what's explanatory about (2') is that it "tells us by what mechanism the causal functions we associate with heat are effected[, and] our knowledge of chemistry and physics makes intelligible how it is that something like the motion of molecules could play the causal role we associate with heat" (1983: 357). He goes on to say that, prior to our discovery of the essential nature of heat, we were already familiar with the causal role that heat played, and that causal role exhausts our notion of it. In other words, we have a pre-theoretic

[3] It's worth noting that David Chalmers is not particularly bothered by the possibility of a sentient country, which he expresses in his 1996 book, *The Conscious Mind*.

understanding of what heat does, and once we have attained an explanation of how that causal role is carried out, there's nothing else that we need to understand about it. It does not seem that we can carry out this same form of explanation with statements like (1). While we certainly feel that the causal role of pain is important to an adequate understanding of what pain is, and certainly it's important to understand the underlying mechanisms that lead to pain, there's an additional feature of pain that is central to it: how it feels. This qualitative character, as Levine calls it, is left entirely unexplained. It is not made clear why C-fibers firing should result in pain feeling the way that it does. As Levine remarks, "there's nothing about C-fiber firing which makes it naturally 'fit' the phenomenal properties [of pain]" (1983: 357). The identity between pain and C-fibers, says Levine, is made into nothing more than a brute fact.

The second major problem posed is that it renders theories like the identity theory deeply uninformative. Brutely identifying pain with brain states or functional states tells us remarkably little. This worry is different from the one presented above. The issue isn't merely that we lack an explanation for the phenomenal feel of mental states; the issue is that the identification of mental states with physical states tells us nothing about "how thickly or thinly to slice our physical kinds when determining which physical state it is identical to" (Levine, 1983: 360). There are at least two issues here. The first is that, presumably, many of the things that feel pain have different physical states from ours. This can have one of two consequences. Either the physical realizers turn out to be highly disjunctive or it turns out that only we can feel pain. Consider the latter. If pain is identical with C-fibers, then things that lack C-fibers cannot feel pain. If they could feel pain without C-fibers, perhaps by having the D-valves of

their hydraulic nervous systems activated, then we would get the absurd result that C-fibers are identical with D-valves (Levine, 1983: 360).[4] Either way, this does not bode well for the identity theory, since neither the consequence that only we feel pain nor the consequence that non-identical things are identical are acceptable. The other option is that the realizers turn out to be highly disjunctive. If that's right, then identifying pain with the firing of C-fibers doesn't tell us anything about what else pain might turn out to be. Indeed, there doesn't seem to be anything that could tell us what else feels pain at all, even though, intuitively, many things feel pain. Secondly, the identification doesn't tell us how much a brain state could change while still counting as pain. Would a minor change in a physical state result in a wildly different phenomenal state like pleasure, or would it still be pain? These brute identifications are deeply uninformative both about why phenomenal states feel as they do and what physical states would need to be like to be phenomenally similar or different from one another. Indeed, the only thing the type identity theory seems to tell us is simply that the mental states we're familiar with are identical with the brain states we've believed them correlated with. Functionalism does not fare much better, telling us only that mental states like pain are identical with some functional state, F, but failing to explain why or tell us what kinds of mental changes we could expect from minor or major functional changes.

According to Levine, physicalist theories of mind cannot successfully solve the major epistemic problems posed by the explanatory gap. The only way out that Levine sees for such

[4] Levine and I both take the D-valve hydraulic nervous system example from David Lewis' "Mad Pain and Martian Pain."

physicalists is to become eliminativists about qualia. This, however, is clearly unacceptable to most. Yet, while Levine believes there is no alternative physicalist solution, I believe there is.

A proper physicalist theory of mind needs a number of things, the first of which is making the connections between the mental and physical intelligible. It needs to remove that feeling of deep mystery as to why this particular physical state feels the way that it does. Such a theory should give us the same satisfaction that we get from statements such as (2'). Furthermore, in order for the theory to be physicalist, Levine claims that a minimal form of reduction is implied. In his words, materialism "implies explanatory reductionism of at least this minimal sort: that for every phenomenon not describable in terms of the fundamental physical magnitudes [...] there is a mechanism that is describable in terms of the fundamental physical magnitudes such that occurrences of the former are intelligible in terms of occurrences of the latter" (1983: 358-9). In other words, there must be some physical mechanism that makes clear, for instance, what makes this particular sensation pain, and that sensation a more intense pain, and this other sensation pleasure. While the theory needn't provide all of the details, there should at least be an abstract understanding as to how such questions are settled. There ought to be symmetry between the phenomenal and physical. This minimal reduction doesn't commit one to a stronger type of reduction such that physics is all there is or that other fields lack autonomy, but it should do away with the kind of brute identifications to which the identity theory commits the field of psychology. Finally, the type of explanation requested should also make sense of when further explanation is unnecessary or inappropriate (Levine, 1983: 358). For instance, an explanation is clearly required for the workings of gravity, but it

would be inappropriate to ask for an explanation of the gravitational constant. Presumably, some facts will be brute, and these facts should be lower-level, fundamental facts. Surely, we don't feel that there's any need for the gravitational constant to be explained, and similarly we should be given rules for when enough explanation has been provided for phenomenal facts. The identity theory clearly fails to attain this.

I argue that none of the major physicalist theories of mind on offer are equipped to deal with the explanatory gap, which I consider to be a deeply concerning problem for said theories. However, I do not agree with Levine that there are no physicalist theories available capable of doing so. I believe that at least some panpsychist theories can appropriately be called physicalist and are prepared to meet the task of closing the explanatory gap in a satisfying manner. The primary task of this dissertation is precisely to articulate one such panpsychist theory.

Before going any further, though, there is one possibility that I wish to dismiss. Perhaps, one might think, it isn't that we've yet to find the right theory. Perhaps it just simply isn't possible to find a theory that provides this satisfying kind of explanation because the relationship between mind and body is beyond human comprehension. I believe this is false, but before continuing, I want to give serious consideration to this possibility as expressed by Colin McGinn. Afterward, I will provide a roadmap of how this dissertation will progress.

Section B: Colin McGinn

Maybe the relationship between mind and body is something that we are simply incapable of wrapping our heads around—perhaps the explanatory gap is an unbridgeable chasm. This is

the view endorsed by McGinn. McGinn rightly claims that we cannot reasonably expect that we will know the answer to every question—this would be the height of human arrogance. Surely, the universe must contain some mysteries that are beyond our comprehension. Finding the elusive mind-body link has been a human project for at least two millenia, and yet it seems clear that we have made absolutely no progress on an answer. We may have made great strides on what Chalmers calls the 'easy problem of consciousness', but none whatsoever on the 'hard problem'. This may be one of the mysteries to which the universe jealously forbids us access: the problem may seem so intractable because it is precisely the kind of thing that lies beyond the scope of human understanding.

McGinn characterizes this inaccessibility in terms of what he calls 'cognitive closure'; roughly, something is cognitively closed to us when our cognitive capacities are limited such that we are incapable of grasping it. That some things in the world will be cognitively closed off to us should strike us as unsurprising. Indeed, there are some clear cases McGinn picks out of commonplace cognitive closure with which we are already familiar. For instance, there is much that is presumably cognitively closed off to rats that would be perfectly within the grasp of chimps. Abstract concepts seem cognitively closed off to young children and accessible to us. Whether something is cognitively closed off to a mind depends on the type of mind that it is. Formally, "a type of mind *M* is cognitively closed with respect to a property *P* (or theory *T*) if and only if the concept-forming procedures at *M*'s disposal cannot extend to a grasp of *P* (or an understanding of *T*)" (McGinn, 1989: 350). Abstracta are beyond the capacities of chimps and infants, and much of what lies within the reach of the chimp lies without the reach of the

rat. This is far from peculiar. That the mind-body link is cognitively closed off to beings like us, then, is at least a live possibility.

In an effort provide an intuitive instance of what it would mean for certain concepts to be cognitively closed to us, McGinn humors David Hume and resuscitates his theory of mind, asking us to envision a being with a true 'Humean mind'. Hume conceived of the human mind as operating entirely on the basis of sense impressions. The concepts that such a Humean mind would form would have to be entirely derivative of sense data. As a result, "the concept-forming system [of such a mind] cannot transcend what can be perceptually presented to the subject" (McGinn, 1989: 351). The Humean mind would be incapable of conceiving of the properties of unobservable entities such as atoms and would be unable to represent those properties. In more general terms, the properties of unobservables would be cognitively closed to the Humean mind. Yet, this doesn't mean that the Humean mind would be incapable of noticing that there's something deeply mysterious about the world that it cannot make sense of. This mind would still be aware of the macroscopic phenomena for which these unobservable entities are responsible. As such, the Humean mind would appreciate that there's a problem in need of an answer in making sense of the world. However, without the proper concept-forming system, the solution would remain forever beyond its reach.

McGinn wants us to believe that the same thing is going on in regards to the relation between the mind and body. We can appreciate that there's a problem, since we experience minds and see brains, and we can further appreciate that the two are intimately tied, but because of our cognitive limitations, the solution to the problem lies beyond our grasp. To be

clear, McGinn isn't interested in establishing the mere possibility that the mind-body link is beyond our reach; he believes that the project is utterly hopeless, and he insists he can prove it. McGinn sees the relation between the mind and body as some property of the brain that makes sense of how it gives rise to consciousness. It is this property *P* that is definitively beyond our understanding. Let us now take a close look at how McGinn establishes the futility of the project at hand.

First, McGinn takes it as obvious that there is some property of the brain that makes sense of the existence of minds. In his words, "resolutely shunning the supernatural, I think it is undeniable that it must be in virtue of *some* natural property of the brain that organisms are conscious" (McGinn, 1989: 353). This natural property must be explainable. Now, he doesn't mean explainable *by us*, but rather explainable in principle. He compares consciousness to the emergence of life. Rightly insisting that being alive is a natural property of bundles of matter, it is subject to natural explanation; well, consciousness is also a biological property, therefore it, too, must be subject to naturalistic explanation "whether or not human beings are capable of arriving at this explanation" (McGinn, 1989: 353). He puts the matter a tad more formally: "Let us say that there exists some property *P*, which fully explains the dependence of conscious states on brain states. If we knew [theory] *T*, then we would have a constructive solution to the mind-body problem. The question then is whether we can ever come to know *T* and grasp the nature of *P*" (McGinn, 1989: 353). The answer from McGinn is a resolute 'no'.

The fact that we have made no meaningful progress over the last few millennia he takes as suggestive. It might be the case, he claims, that we have yet to be gifted the "Einstein-like

genius" who will phrase the problem differently or otherwise challenge the way we've been approaching the issue and will provide us with the solution we've been looking for (McGinn, 1989: 354). This, however, he believes is unlikely. We find the problem so baffling that we should be open to the possibility that there is no solution that we'll come to. Indeed, there are only two ways we might try to find the ever-elusive property P, he says.

The first way we might try to find P is by looking directly at our consciousness. The tool we use is, of course, introspection. He believes this cannot work. Through introspection, we can become intimately familiar with all of the dazzling peculiarities of mentality, but this is only one side of the equation: "we have direct cognitive access to one term of the mind-body relation, but we do not have [introspective] access to the nature of the link" (McGinn, 1989: 354). We can introspect all we like, but the process of introspection, he rightly insists, does not reveal our mentality as being dependent on the brain. We have a second option, though, which is to study the brain itself. This, too, is doomed to fail. The problem is that our minds represent the material world spatially. Indeed, we form our physical concepts in spatial terms. When we look at brains, what we see are wrinkly lumps of gray matter extended in space, but consciousness just isn't a spatial property, and whatever it is that links consciousness to brains will not be sensible in spatial terms (McGinn, 1989: 357). If our minds cannot operate outside of a spatial framework, and if it turns out that *P* is non-spatial, then we simply will not grasp *P*.

McGinn submits that the missing link between the mind and body is actually very likely to be simple. This gives all the more reason to believe that what we're dealing with here is a problem of cognitive closure. Consciousness, McGinn says, comes along in the evolutionary

chain far before things like language. This suggests that whatever it takes to bring about consciousness from brains must be easier to accomplish than what it takes to bring about linguistic systems. And yet we've made incredible progress in our understanding of the latter. Apparently, our inability to make sense of the missing link has nothing to do with its being particularly complex or obscure, it's simply cognitively closed to human beings. There may very well be other beings out there that have the concept-forming apparatus necessary to make sense of the problem easily, and their minds are likely to be quite different from our own.

McGinn makes two mistakes. Concerning the simplicity of *P*, I believe that his evolutionary point is misleading. Where something shows up on the evolutionary line needn't be indicative of how difficult it is to grasp conceptually. His claim is that the fact that consciousness precedes language in the evolutionary lineage shows that it was easier for nature to bring consciousness about than it was for it to bring language about. So, it should be easier to grasp the mysteries of consciousness than it is to grasp those of language. But this inferential leap is a bad one. It seems that McGinn thinks that the fact that we've made progress on language and not consciousness is supposed to suggest, given the evolutionary record, that the latter is cognitively closed to us. If it weren't, then surely we would have made more progress on consciousness than language, since it precedes language. But where something lands on the evolutionary lineage doesn't say anything about how difficult it is to explain it. Making life from non-life was the very first evolutionary step, presumably preceding consciousness. Nonetheless, we are much closer to understanding language than we are to understanding

abiogenesis.[5] Yet no one would suggest on the basis of where language and abiogenesis land on the evolutionary timeline that the answer to either is cognitively closed to us. That we've made progress on language but not on consciousness is no evidence at all that the latter is beyond our reach.

His second mistake is in presenting our options in searching for an answer as a dichotomy between introspection and neuroscience. There is a third horn, and it's one that McGinn dismisses too quickly. Perhaps we've been looking at the problem the wrong way, and I'm skeptical that we'll require a new Einstein in order to reframe things. The core of the problem posed by the explanatory gap is that we cannot imagine what it is about the physical that could possibly settle that mental phenomena are as they are. In this dissertation, I hope to prove at the very least that the answer is conceivable: there is one way of understanding the world that can shine light on why physical states bring about the mental states that they do, though it will require rejecting some tenaciously held assumptions and making way for frightening new ones. Either way, that an answer is beyond our reach is not something we can come to know simply because of the difficulty of the problem at hand. McGinn's claim that the explanatory gap cannot be bridged is a possibility, but until the intellectual well runs dry, we must continue our search for the answer.

I now wish to turn my attention to Frank Jackson's famous Mary's Room thought experiment to elucidate exactly what it is that feels so problematic about the explanatory gap. Jackson's thought experiment threatened physicalism by an indirect appeal to the gap, and so I

[5] The issue of the difficulty of abiogenesis is well-known, and while some theoretical progress has been made, there is little consensus. E.g., see Peretó (2005), Scharf et al. (2015).

will look at some of the responses it prompted. Those responses, I argue, leave the explanatory gap wide open. The problem, I believe, is deeply pervasive, and so I then turn my attention to physicalist theories of mind more generally. I will consider their strongest formulations and their most powerful objections, each of which has successfully shown the gap to remain. Finally, I end the chapter by picking out precisely what these theories are missing that results in their inability to solve the problem posed by the explanatory gap. This allows us to figure out which qualities a theory of mind would need in order to provide us with a satisfying explanation.

Section C: Mary's Room

Jackson considers himself, at least at the time of writing "Epiphenomenal Qualia," a qualiophile: someone who takes qualia very seriously. I myself am in the same camp, and, indeed, most philosophers are likely to believe that qualia must be accounted for in some way. At the same time, most philosophers consider themselves physicalists. Jackson, however, believes that the physicalist doctrine is doomed to leave qualia out of the picture. Physicalism, he argues, simply doesn't have room for qualia. His famous Mary's Room thought experiment and his oft-neglected Fred case are both provocative, and Jackson believes they reveal an irreparable flaw in physicalist thinking. According to Jackson, there is "nothing you could tell of a physical sort [that] captures the smell of a rose [...,] therefore, physicalism is false" (1982: 127). I disagree, along with Terence Horgan, that physicalism perishes at the hands of his clever thought experiments: Jackson is equivocating two types of physicalist information, and it does

not follow from Mary's new knowledge that physicalism is false (1984). He does, nonetheless, bring to light a common flaw in physicalist thinking that proves troubling, so I will follow him for now. Ultimately, I will defend physicalism with the help of Earl Conee and Horgan.

Jackson construes physicalism as the view that the only facts that exist are physical facts, and physicalists are by and large happy to agree with this characterization. When the time comes to deal with the pesky question about qualia, many physicalists find themselves pressured into claiming that qualia have already been accounted for by their physicalist theories. Think back to the identity theorist's position: mental phenomena just are physical phenomena, and that's the end of the story. This is, of course, deeply unsatisfying, but Jackson's project makes salient why this kind of response is so bothersome. Let us now turn our attention to the thought experiments themselves.

Jackson's most popular thought experiment concerns Mary, a color scientist who knows all of the physical facts about color. She knows to which frequencies the terms 'red', 'green', and so on refer, and she knows precisely how those frequencies interact with our retinas ultimately to produce certain neuronal excitations. Indeed, every single physical fact about color that could be known *is* known by Mary. However, Mary was born and raised in a black-and-white room. She learned everything she knows through black-and-white textbooks and a black-and-white monitor. Note, Mary's brain is perfectly normal—were she to see color, she would still perceive it as one normally does. It's just that she's never had the pleasure. But here comes the end to Mary's colorless days: Mary is finally allowed to step outside of her black-and-white chamber, and upon doing so, she is presented with a ripe tomato. And so the

question goes: upon seeing the tomato, does Mary learn anything new? The intuitive answer is: yes! Mary now knows *what red looks like*.[6] This is precisely the response Jackson finds problematic for the physicalist. Recall, the only facts that exist are the physical facts, and Mary already knew all of those before leaving her chamber. So, if we admit that Mary has learned something new, then Mary must have learned a non-physical fact. But this is just to say that physicalism is false.

To flag and temporarily put aside an important concern one might have at this point: one could claim that the intuition appealed to here is misguided. Given that we've allowed Mary to possess *all* of the physical facts before granting her access to the outside world, perhaps we have already granted Mary color experience. Perhaps upon leaving her chamber, Mary won't learn anything new after all. It may be that our intuition is misguided, and the physical facts alone really do suffice for phenomenal knowledge. This possibility is defended by Daniel Dennett, and it's one I will delve into quite soon. I will argue that Dennett makes a fatal mistake and accidentally ends up agreeing with Jackson's central point. The intuition driving Jackson's thought experiment is so strong, Dennett himself unwittingly gives into it. However, for now, let's return to Jackson's thought experiments, this time taking a close look at Fred.

Fred is a man with exceptional color vision. Where we would normally see only a single shade of red, Fred can apparently see two colors. To prove this, we have Fred sort out apples of what is apparently the same shade of red, and he always successfully sorts them into the same two groups, even though we keep shuffling them up when he isn't looking. Indeed, Fred never

[6] She would probably have no way of knowing that *this* color is the one that the term 'red' denotes, though she may know enough about ripe tomatoes to know that she's very likely experiencing redness rather than blueness.

makes the mistake of misplacing so much as a single apple, and when asked, he claims that the two colors he sees aren't merely different shades of a single color, but actually entirely different hues. Furthermore, "an investigation of the physiological basis of Fred's exceptional ability reveals that Fred's optical system is able to separate out two groups of wavelengths in the red spectrum as sharply as we are able to sort out yellow from blue. I think that we should admit that Fred can see, really see, at least one more colour than we can..." (Jackson, 1982: 128). In the case of Fred, the rest of us play the role of Mary. We have always been locked up in our own chamber where the colors that Fred can see do not exist. I find the case of Fred more compelling than Mary's—it drives home the intuition more powerfully: it seems obvious that there is nothing we could do that would allow us to see what Fred sees. We can find the minor differences in the relevant electromagnetic frequencies, study precisely what Fred's brain does when in that particular state, and so on for the rest of the physical facts, and yet we still would have no idea what Fred is experiencing. After poking around in Fred's physiology and possessing "all the physical information we could desire about his body and brain," it would still be the case that "there is more to know than all that" (Jackson, 1982: 129). Thus, it seems that the physicalist picture necessarily leaves something out.

The problem that Jackson is indirectly appealing to is the explanatory gap between mind and body. The physical information possessed by Mary does not suffice to settle the phenomenal facts. It seems that even given Mary's vast knowledge, what it is like to see red could have been anything. The physical facts leave the phenomenal facts completely open. Similarly, it seems that Fred might be seeing anything. We have no idea what Fred's color

experience is like, and no amount of physical information will prove illuminating. “If Physicalism were true, enough physical information about Fred would obviate any need to extrapolate or to perform special feats of imagination or understanding in order to know all about his special color experience. The information would already be in our possession. But it clearly isn’t” (Jackson, 1982: 132).

If the physicalist wishes to resist Jackson’s conclusion, there is only one option available: one must deny that phenomenal facts are non-physical facts. There are a few ways one may attempt to achieve this, and we will be investigating them shortly. One way would be to claim that phenomenal facts actually do turn out to be physical. Given that facts can be written in textbooks and phenomenal experiences cannot, it must turn out that phenomenal facts must be somehow derivable from standardly non-phenomenal physical ones, one might argue. In other words, perhaps Mary can infer the redness of the tomato from what she knows about her neurological dispositions or otherwise, on the basis of purely non-phenomenal physical information, come to possess phenomenal physical information. Alternatively, one might claim that phenomenal knowledge is simply not factive. If so, then there would be no such phenomenal, non-physical facts to threaten physicalism, as the phenomenal may not deal in facts at all. Phenomenology survives, but it is some other kind of beast that promises no harm to the physicalist doctrine. If, however, Jackson is right, and phenomenal facts are non-physical and, indeed, facts, then physicalism is in trouble. Naturally, a number of philosophers have attempted to show that Jackson is wrong. We will now transition to considering some of these replies. Specifically, we’ll look at powerful responses that have been

provided by Daniel Dennett, David Lewis, Terence Horgan, and Earl Conee. I will argue that Dennett's response accidentally concedes the point to Jackson, Lewis' provides an inadequate explanation of what Mary is doing, and Horgan's and Conee's leaves the central problem Jackson appeals to—the explanatory gap—untouched. I will, however, agree with Conee, Horgan, and Lewis that phenomenal knowledge is not knowledge of facts. Afterwards, we will move onto the physicalist theories of mind we have been dealing with indirectly in the hopes of uncovering precisely where they err.

Section D: Dennett's RoboMary

Daniel Dennett, in "What RoboMary Knows," takes issue with Jackson's thought experiment. According to Dennett, whether Mary learns anything new upon leaving her room could go either way. Indeed, he provides an alternative version of Mary's room in which, upon leaving her black and white chambers, she's handed a blue banana in an attempt to fool her, but she surprises her pranksters by exclaiming that she's well aware that bananas are yellow and yet this one is blue. In this scenario, Mary is completely unsurprised by what blue looks like. The reason, claims Dennett, is precisely how much physical information she was already privy to prior to leaving the room—this alternative to Mary's room is equally plausible, it's just hard for us to conceive of the vastness of Mary's knowledge prior to leaving the room. It's hard to for us imagine what it would even mean for someone to possess *all* of the physical facts about color. The fact is, the intuition that we rely on in claiming that Mary learns something new is not to be trusted. We should not just assume that it is impossible to *figure out* what color experience is

like. We might, for instance, discover some ingenious proof or inferential technique that could get the job done.

Of course, it seems pretty clear to us that no amount of information of a physical kind could possibly lead to phenomenal information—this is the very intuition that leads us to believe that one could never explain to a blind person what color is like. But perhaps, Dennett claims, given the amount of information Mary has access to, she actually could deduce the phenomenology of color from a 3000-step proof. The point Dennett wants to make is that the Mary's Room thought experiment as provided by Jackson is too quick. However, Dennett has a new version that he believes makes clear how someone who possesses the kind of information Mary has access to could actually figure out what color is like from the confines of her room. If he's right—if Dennett succeeds in showing us a way for someone to discover what the phenomenal facts are given only the physical ones—then Jackson's argument fails: it turns out the physical information really is enough to settle the phenomenal information. Jackson simply told the wrong version of Mary's room. Unfortunately for Dennett, far from proving his point, his version of Mary's Room serves only Jackson's ends. We'll see how shortly, but first, let us turn to Dennett's more intricate and certainly more careful version: RoboMary's Room.

Dennett's RoboMary thought experiment is specifically designed to make clear how a nearly-omniscient being (at least omniscient when it comes to the physical color-facts) could arrive at phenomenal facts from physical facts. Before we delve into the thought experiment itself, there are a few important preliminaries. First, we need to get clear on precisely what Dennett needs to prove. What Dennett will most stress in this thought experiment is the

wealth of information his color scientist has available. It is from this information that she must come to learn the color information that Jackson believes to be eternally elusive and which Dennett believes is within her grasp, claiming: "Mary had figured out, using her vast knowledge of color science, exactly what it would be like for her to see something red, something yellow, something blue in advance of having those experiences" (2005: 106). To reiterate, Jackson's main claim is that from the physical facts alone, it is impossible for Mary to learn the phenomenal facts, which he takes to mean that those phenomenal facts lie outside the realm of physicalism. Thus, in order for Dennett to be successful, he must show how Mary's vast knowledge (or, in Dennett's case, RoboMary's) does indeed grant her access to those elusive phenomenal facts. As he reminds us time and again, "Mary knows *everything* about color that can be learned by physical science..." (2005: 115). Dennett provides us with two versions of his thought experiment. I argue that both fail for the same reason.

Meet RoboMary. RoboMary is a robotic version of Mary. In order to avoid superficial objections, Dennett asks that we set aside any concerns about sentient robots. Although he is adamant in his paper that sentient robots are very much possible, allowing RoboMary to replace Mary is dialectically beneficial regardless of where we fall on the conscious AI debate, as we have a greater understanding of the inner workings of a robot's machinery than we do of neuroscience, and Dennett heavily relies on this machinery in making his points. Now, RoboMary needn't be confined to a black and white room. She is a robot who has been built with a set of camera-eyes that provide her with only black-and-white visual data. At the end of her career in becoming omniscient in regards to physical color facts, Mary's black-and-white

camera eyes will be replaced with color-receptive ones. We will then discover whether RoboMary learns anything new upon first experiencing color. In order to have such color experiences, of course, RoboMary's brain (alternatively, her circuitry) is perfectly capable of perceiving color. This bit is vital. Given that Mary's visual system is intact, then, in Dennett's words: "she already has 'in there' everything she needs to experience color; it just hasn't been stimulated" (2005: 118). Note: what's at issue cannot simply be whether RoboMary can find some way of activating the right neurons or circuitry. Such a task could be accomplished by significantly less-informed persons, either by directly seeing tomatoes or probing their brains in the right way. These things are actions, not facts. What matters, once again, is whether RoboMary, from her vast wealth of knowledge, can come to know the phenomenal facts from the physical facts alone—this is what Dennett must establish. Dennett himself seems to acknowledge this, as he says that "what *is* worth discussing" is the scenario in which "Mary puts all of her scientific knowledge of color to use and *figures out* exactly what it is like to see red (and green, and blue) and hence is not the least bit surprised when she sees her first rose" (2005: 122). The key phrase here is 'figures out', in Dennett's own words. Once she has all of the physical facts necessary, she must come to learn from those facts what the phenomenal facts are. To assist her, RoboMary has access to her companions, the Mark19s, who are identical to RoboMary except for their fully-functional, color-capable camera eyes.

In the first version of Dennett's thought experiment, RoboMary, through studying her robot companions, the Mark19s, learns how they encode color information and save that information in their registers. After learning the precise process by which the Mark19s do

this—by studying their robo-physiology, their dispositional states when it comes to perceiving colors, etc.—RoboMary writes a program that takes the information her own black-and-white camera eyes receive and converts that information into color information. This program, the use of which Dennett calls 'imagining', simply colorizes the input from her eyes (2005: 124). The process wouldn't be perfect at first, but, since RoboMary knows everything physical about color, she knows precisely how she *should* react to certain colors, and she knows how her companions *do* react, and with this comparative information, she can make the requisite tweaks in her 'imagine' program that will then superimpose the appropriate colored pixels on the black-and-white images she receives from her cameras. With enough tweaking, she becomes capable of 'imagining' what the correct colors of everything are, and upon receiving her color-capable camera eyes, she is not surprised by what she sees.

As I see it, this is straightforwardly not an instance of figuring out what color is like. This is the kind of trick that many philosophers have tried to proof Jackson's original thought experiment against, such as the possibility of dreaming colors or generating phosphenes by pressing up against one's eyes. Dennett, though largely unsympathetic, is receptive to this objection, and so he constructs an alternative. In this first version, it's clear that RoboMary did not figure out the color facts from the physical facts—she merely figured out some way of activating the right circuits so that she would experience color prior to receiving her better camera eyes. There's a deeper problem with this version of the thought experiment, but we'll come back to this right after presenting the next version.

In the second version of the thought experiment, RoboMary is barred from creating any such programs. Furthermore, the color registers in her brain are locked to grayscale values until she gets her new color-capable camera eyes. RoboMary then decides to get creative. She creates a model of herself that can perceive color. She studies this model: how it is disposed to react to color, how its wiring works, etc., and she does the same with herself. Because of her grayscale limitations, RoboMary's dispositions are different from those of her model. Thus, when she sees an apple, she is in 'state A', and when her model sees an apple, the model is in 'state B'. To ensure complete loyalty to Dennett, what happens next is in his words:

> RoboMary notes all the differences between state A, the state she was thrown into by her locked color system, and state B, the state she would have been thrown into had her color system not been locked, and—being such a clever, indefatigable and nearly omniscient being—makes all the necessary adjustments and *puts herself into state B*. [...] But now she can know just what it is like for her to see a red tomato, because she has managed to put herself into just such a dispositional state... (2005: 127-8).

Dennett doesn't elucidate what it means when RoboMary 'puts herself into state B', but we can imagine any number of possibilities. Considering Dennett's commitments elsewhere, I imagine that putting herself into this state has something to do with changing her circuitry so that her dispositions line up exactly with those of her model. This, Dennett believes, is enough to ensure that RoboMary is capable of seeing color. This is also, however, where Dennett makes a fatal mistake.

To understand the mistake, let's take a look at what RoboMary is doing. We'll go step-by-step. First, RoboMary learns all of the physical facts. This happens when she studies her model and herself and presumably everything else that's physically relevant. Second, once

she has acquired all of the physical facts, she now knows precisely what state she would need to be in—how her circuits would need to light up—in order to perceive color as her model does. Third, and last, she then *puts herself into that state*. And so Dennett gives the game away to Jackson. Notice how there's a point between steps two and three where RoboMary has all of the physical knowledge, but she has yet to put herself into state B. This is the point in the timeline where she possesses all of the physical facts, but she has yet to activate the right circuits. Nonetheless, she *knows* which circuits need to light up. She knows her model's dispositions. She even knows what she needs to do in order to get them to light up. Yet, until she actually occupies state B, she has no idea what red looks like. Indeed, at that particular point in time between steps two and three, what color looks like is a big question mark for RoboMary. It appears that all of the physical information in the world has proven insufficient for RoboMary to *figure out* what the phenomenal facts are. It isn't until RoboMary *does* something that she can come to experience color. Dennett characterizes this type of knowledge as merely knowledge about physiology, claiming that it is "simply a way of dramatizing the immense knowledge of color 'physiology' that RoboMary" enjoys (2005: 124). But that cannot be right, as physiological knowledge can be possessed without thereby changing one's physiology, and it seems that the former is insufficient. She needs to act on her knowledge in order to attain color experience. But this is just to concede that there's something extra that RoboMary and Mary need to do in addition to possessing all of the physical facts: they have to actually experience the colors. This is precisely the problem that occurs in the first version Dennett provides. In both versions, RoboMary obtains all of the physical information, but she

has no idea what color is like until she experiences that color one way or another. Unless the relevant neurons or circuits are activated, the Marys are in the dark.

Let's condense the issue. Reconsider Jackson's version. The objective was not whether Mary can somehow experience color while within the room by any means necessary. If that were the objective, we might as well just give Mary a window. The objective is to see whether Mary, on the basis of physical *facts* can come to learn phenomenal facts. Both of Dennett's versions are just an instance of placing a window in Mary's room. This doesn't get at the heart of the issue. The only reason RoboMary is unsurprised upon attaining her color cameras is because she peeked out the window, and being omniscient of all of the physical color facts is quite unnecessary for this. Dennett believes that all that's required is that Mary (or RoboMary) learn all of the appropriate dispositions and then place herself in them. This is clear in both of his versions, since he seems to think that upon knowing the exact state she'd need to be in, she would have narrowed down what the experience will be like to just one. But this is not the case. Even when she has complete knowledge of what her dispositional character will be like upon seeing the color we call 'red', she still has not grasped what such a color will be like.

The explanatory gap is just as present in Dennett's versions. In order for Dennett to have successfully shown that Jackson's thought experiment ends the wrong way, he would have had to show that RoboMary's physical information proves sufficient. Yet, once again, the physical facts appear to leave the phenomenal facts wide open. Dennett's response to Jackson still leaves us scratching our heads: RoboMary knew *everything* there was to know physically, and yet, until she actually experienced the colors herself, everything she knew failed to settle the

phenomenal. Actions are not facts, and yet it was actions that were necessary for her to leave the confines of her room unsurprised.

We have now considered one potential response to Jackson: Mary's Room does not prove that physicalism is false because Mary does know what color is like before leaving her room. We've seen this response fails, so now we turn our attention to a different kind of response as defended by David Lewis: Mary *does* learn something new upon leaving her room, but what she learns is no threat to physicalism. She has simply learned a new skill.

Section E: David Lewis' Knowledge How

As I have stated previously, it appears that no amount of physical information could amount to phenomenal information. Lewis agrees, stating that it "won't help at all to take lessons on the composition of skunk scent or Vegemite, the physiology of the nostrils or the taste-buds, and the neurophysiology of the sensory nerves and the brain," for none of that will tell us what the relevant experiences are like (1990: 500). To learn what Vegemite tastes like, we'll have to try some actual Vegemite. Stated more explicitly, Lewis believes that perhaps the only way to learn what an experience is like is to have the experience. Crucially, he says that "there is a change that takes place in you when you have the experience and thereby come to know what it's like. Perhaps the exact same change could in principle be produced in you by precise neurosurgery..." (Lewis, 1990: 500). This may sound familiar, as it's a point that I raised above in response to Dennett; namely, that information proved to be insufficient and RoboMary had to have the right circuits light up in order to experience red. It seems that what matters the

most is that the right stuff gets activated. The solution that Lewis proposes to the problem posed by Mary's Room will have to deal with similar concerns.

First, a quick note: the world very well could have been such that physical information brought phenomenal information along with it as a mere matter of contingent fact. Lewis shares the same view, noting that "just as we can imagine that a spell might produce the same change as a smell, so likewise can we imagine that science lessons might cause that same change. [...] There might have been a causal mechanism that transforms science lessons into whatever it is that experience gives us" (1990: 500-1). It's important to point out that in such a world, the explanatory gap would pose all of the same problems, for nothing would explain why *this* explanation of the neurological process involved in the taste of Vegemite results in *this* particular taste (which just so happens to be the taste of Vegemite) as opposed to some other taste. Even if such science lessons really did exist, they would still largely leave us in the dark. If such a lesson taught us what it's like to be a bat, we still would have no idea that what we learned (and thereby experienced) really is the experience of being a bat as opposed to some other non-bat-like experience.

Lessons may not cut it, but Lewis argues, as we'll get to momentarily, that phenomenal information is the possession of a new skill. If that's right, then the apparent ineffability of phenomenology will be accounted for and we will save physicalism in one fell swoop. But, in order for Lewis to explain away the problem of Mary's Room successfully, he must do at least two things. First, he must show that know-how is sufficient for phenomenal knowledge. I will argue that it is not, and it is in part because the explanatory gap rears its ugly head once again.

Second, he must show that know-how is necessary for phenomenal knowledge, and this, too, I argue is unsuccessful.

Let's begin by fleshing out Lewis' view. In his words: "If you have a new experience, you gain abilities to remember and to imagine. After you taste Vegemite, and you learn what it's like, you can afterward remember the experience you had. By remembering how it once was, you can afterward imagine such an experience" (Lewis, 1990: 515). In this sense, learning the taste of Vegemite is the same as learning how to ride a bike. It isn't a new form of knowledge that exists in a non-physical realm; it's just knowledge that we cannot impart through mere lessons. Telling someone what they need to do to ride a bike does very little to help one ride one. One needs the have the experience oneself to truly know. To be clear, the thesis is that *all there is* to learning phenomenal information is the acquisition of some skills. Perhaps his clearest iteration of the thesis is this: "The Ability Hypothesis says that knowing what an experience is like just *is* the possession of these abilities to remember, imagine, and recognize. It isn't the possession of any kind of information, ordinary or peculiar. [...] It isn't knowing-that. It's knowing-how" (Lewis 1990: 516).

That knowing what it's like to experience red is know-how is, at the face of it, plausible. Lewis' claim is about what it means to possess a type of knowledge. Phenomenal knowledge, on Lewis' view, is the possession of skill-knowledge. Once we possess that knowledge, we can choose to exercise it. On command, I can see red if I so choose. This is thanks to my imaginative skills. I also needn't be exercising this skill in order for it to be the case that I know what red looks like. I do know what red looks like, and yet I am currently neither remembering

it, imagining it, nor recognizing it within my field of view. But we must remember why Lewis offers this proposal. It is in response to Jackson's Mary's Room. The central issue with the thought experiment is the existence of the explanatory gap. In order for a response to Jackson to be successful, it must do away with the gap. Lewis hopes that by identifying phenomenal knowledge with know-how the gap can be closed. But to show that the gap has been closed and Jackson's problem solved, Lewis needs to do more than identify phenomenal knowledge with a type of know-how: he needs to show that the exercise of this skill explains why Mary sees what she sees. When one rides a bike, one exercises one's ability to do so, and it is clear from the physiological happenings in conjunction with the neighboring physical happenings how it all adds up to bike-riding. To close the gap, it must be equally clear how Mary's neurophysiological exertions add up to red-seeing. Unfortunately, Lewis' theory promises no way forward, and the fact that his theory does not close the gap undermines his claims, as we'll see.

What Lewis says seems pretty straightforward, but he glosses over precisely what it would mean for phenomenal knowledge to be a skill. Once we get clear on what this entails, the problems become apparent. His Ability Hypothesis claims that learning phenomenal knowledge is nothing more than skill acquisition. Well, which skills? The ones he mentions are memory, imagination, and recognition. Let's consider what these skills entail. In the same way that riding a bike is a physical skill to move one's muscles in such a way that one can maintain balance, move one's legs in a rotating motion, etc., the skills of memory, imagination, and recognition are bound to be neurological skills. For simplicity, we'll discuss only imagination,

as the other skills are similar enough that what we have to say here will apply to them as well. To have the ability to imagine is to have the ability to fire off certain neurons—when we imagine, there is something neurological that we are doing. Recall Dennett. RoboMary, prior to leaving her room, learned how to put herself in state B. We can call that process imagination, if we'd like. Well, for us to learn how to imagine red, the skill we acquire is how to set off some chain of neurons—say Neuronal Chain R—that ultimately activates the relevant cluster of neurons—say Neuronal Cluster R—which results in our experience of redness. Now, I believe Lewis to be right that when we imagine something, we are applying a skill, and it's a skill we acquire upon experiencing the appropriate phenomenon. Hence, upon seeing red, presumably some neurons go off that provide us with the experience, and we then learn how to reactivate the relevant neurons. Put succinctly, then, the skills that we acquire when it comes to phenomenal knowledge are all about learning how to activate the appropriate neurons. This is what Lewis' view amounts to.

I will now argue that the possession of these neurological skills is neither necessary nor sufficient for the possession of phenomenal knowledge. Let's begin by considering whether possessing the skill of imagination is sufficient for possessing knowledge of what redness is like. Let's construct a precise, if inaccurate, neural story that the Ability Hypothesis would require to work. Suppose that we experience red when Neuronal Cluster R gets activated. Neuronal Cluster R appears to be something that we learn to activate by stimulating Neuronal Chain R. We learn to stimulate Neuronal Chain R by seeing red for the first time.[7] What we would learn

[7] While oversimplified, this seems to be how things work. Within our brains there appear to be some neurons that get stimulated to give us the experience of redness, but we don't activate those directly. Our eyes, optic nerve, and

to do on Lewis' view, then, would be learning how to activate Neuronal Chain R, which would in turn activate Neuronal Cluster R, which then somehow gives rise to our experience of redness. Neuronal Cluster R is our neurological bike, and activating Neuronal Chain R is our ability to ride it. So long as we know how to activate Neuronal Chain R, Lewis would say that we know how to see red. But knowing how to activate Neuronal Chain R does not count as knowing what redness is like. Suppose I swapped out, through complex neurosurgery, Neuronal Cluster R and replaced it with Neuronal Cluster B such that now, when you activate Neuronal Chain R, you experience blueness instead of redness. You still know precisely what redness looks like, but for some reason whenever you try to imagine it, only blueness comes up. We can even picture the frustration we would feel at this misfiring of our imagination. We could do the same for recognition, and it would seem as though the world lost its redness. We may not be able to do the same with memory and notice a difference, but this hardly saves Lewis' view.

For it to be true that we know how to ride a bike, knowing how to move our bodies and balance ourselves is sufficient. We could destroy every bike in existence, and it would no less be true that one knows how to ride a bike. We could replace the bike we are riding with a boulder, which would immediately cause us to cease moving our legs, or at least, if we are tenacious enough, to continue rotating them clumsily. But we will not suddenly be boulder-riding, nor will our skill suddenly become a skill to ride boulders. Yet, after the

---

so on serve as Neuronal Chain R, and it is distinct from the bundle of neurons that are directly responsible for the red experience. Indeed, it seems perfectly reasonable that we could bypass Neuronal Chain R and stimulate Neuronal Cluster R directly with some diodes.

neurosurgery, our ability suddenly changes to that of a different color. And we need not restrict ourselves to colors; nothing is stopping us from being truly devious and replacing redness with the taste of Vegemite. Note that we need not even utilize neurosurgery; the explanatory gap takes care of the work for us. We can simply imagine that Mary saw any color at all upon seeing the ripe tomato. We don't actually know what she saw; we know only that she had a new experience. Thus, the skills Lewis is envisioning are not recognition, memory, and imagination *of red* (and so on for the other colors), but rather something more abstract, like general experiential recognition, memory, and imagination. Possessing this skill is not sufficient for the possession of knowledge of what it is like to see red, as its execution need not lead to the experience of redness.

Perhaps I have been unfair. It's possible that I have misdescribed the situation that Lewis had in mind. What if there is no distinction between Neuronal Cluster R and Neuronal Chain R? All it takes to experience red is the stimulation of Neuronal Chain R, one might say. Let's simplify things. We'll nix all talk of Neuronal Chain R. Suppose we can learn how to directly stimulate Neuronal Cluster R. Thus, upon learning how to see red, we've learned how to directly turn Neuronal Cluster R on, hence our experience of red. Would this not then prove that experiencing red is just the acquisition of a new skill, namely, the skill of activating Neuronal Cluster R? If so, then the possession of the skill would inevitably lead to the knowledge of redness; we would not be capable of divorcing the two. Learning *how* would have to be sufficient for our ability to experience redness. However, this possibility is just what Dennett's RoboMary learns to do, and it tells us nothing about why stimulating this particular

neuronal cluster gives us the experience of red. It clearly wouldn't be the *ability* of stimulating Neuronal Cluster R that counts as knowing what the experience is like, as we can easily do that (in theory) by poking the right neuron with a diode—an ability I could possess without ever stimulating Neuronal Cluster R. Consider: it might be the case right now that there is some neuronal cluster in my head that, when stimulated, grants me the experience of ultraviolet. It could further be the case that, right now, I possess the skill of activating it. I have in my hands a button that, when pressed, directly stimulates this cluster. I have yet to press it, however. Surely, I do not possess phenomenal knowledge of ultraviolet. We could even add more buttons so that, in looking around, I can recognize ultraviolet light, and in remembering, my brain will bring forth all of the ultraviolet that had previously gone unnoticed. Say what we will about how cheaply I came to possess these skills, it is nonetheless true that I have the disposition to bring these experiences about, but until they have been brought about, I cannot be said to possess any knowledge whatsoever of what ultraviolet is like. What is relevant to my possessing phenomenal knowledge is my experience of it, not my ability to bring it about. Furthermore, if we believe that stimulating Neuronal Cluster R could have led to any experience whatsoever, then the explanatory gap remains open, and any talk of abilities does nothing to close it.

We have seen that Lewis' Ability Hypothesis fails to supply us with a sufficient condition for experience: the possession of the relevant skills does not suffice for phenomenal knowledge. We now turn to Earl Conee, who supplies us with an objection to Lewis which

effectively shows that the Ability Hypothesis fails to establish the acquisition of skills as a necessary condition for experience.

Conee identifies the usual method employed in defending physicalism against Jackson's knowledge argument. He says, "the main line of critical response proceeds by acknowledging that Jackson has identified knowledge which is not knowledge of physical information, while denying that it is knowledge of non-physical information. It is claimed not to be knowledge of information at all" (Conee, 1994: 136). As we've seen, Lewis does this by making phenomenal knowledge out to be a kind of know-how, claiming that Mary "gains abilities rather than factual knowledge" (Conee, 1994: 138).

According to Conee, however, the Ability Hypothesis requires too much. A person can lack all of the skills required by the Ability Hypothesis and yet experience red no less.

> It requires too much of a person in order to know what an experience is like. [...] To see this, let us suppose that Mary is in the epistemic condition that Jackson ascribes to her during her confinement. She is as well informed about color vision as can be accomplished by lessons in black and white. But suppose too that Mary has no visual imagination. She is unable to visualize anything. [...] Mary is released from her black and white confinement and sees something red for the first time. At that point, while she is intently gazing at the colour of red ripe tomatoes, it is clearly true that she knows what it is like to see something red. [...] Yet, she is unable to imagine anything [...], she is not able to imagine, remember, and recognize the experience, as Lewis' Ability Hypothesis requires in order of her [sic] to know what it is like to see red. [...] Hence, knowing what an experience is like does not imply having any such abilities. (Conee, 1994: 139)

This is clearly right. An individual can lack the skills of recognition, imagination, and memory and still be capable of experiencing redness. That which is most intimately tied to the having of

an experience is the set of neurons that give rise to that experience, and the avenue by which one activates those (be it by memory, imagination, etc.) cannot be the experience itself. Thus, the skills mentioned in the Ability Hypothesis are not necessary for the possession of phenomenal knowledge.

To summarize, Lewis argues that phenomenal knowledge is the possession of a certain set of skills: the ability to remember, imagine, and recognize. As we've seen, these skills would just be ways of activating the neurons that are relevant for the having of an experience. The acquisition of the relevant skills, then, needs to be both necessary and sufficient for the possession of phenomenal knowledge. But we have seen that this is not the case. One can possess the skill and lack the experience and one can also possess the experience itself without having any of the relevant skills. Mary can have red-inducing buttons without ever having pressed them, or, alternatively, she could stare at red ripe tomatoes and lack, due to some deficiency, all of the skills mentioned above. Therefore, it cannot be the case that phenomenal knowledge is know-how. Conee offers an alternative view of what phenomenal knowledge is that promises to keep physicalism safe, or, at least, out of the crosshairs of the knowledge argument: to know a phenomenon is to be directly acquainted with it. It is to this view that we now shift our attention.

Section F: Conee's Knowledge by Acquaintance

Conee argues that phenomenal knowledge is neither factual nor is it the possession of any skills. Instead, phenomenal knowledge falls into a third category. According to Conee,

phenomenal "knowledge consists in acquaintance with the experience [and] does not require having either information or abilities. Acquaintance constitutes a third category of knowledge, irreducible to factual knowledge or knowing how. Knowledge by acquaintance of an experience requires only a maximally direct cognitive relation to the experience" (1994: 140). In other words, acquaintance does not require any kind of information, it is simply a direct relation to the thing with which one is acquainted. Acquaintance with a property, he points out, is analogous to acquaintance with a city; one comes to know a city only by becoming acquainted with it (Conee, 1994: 140).

Now, Conee claims that "learning what an experience is like is identical to becoming acquainted with the experience" (1994: 140). Additionally, he says, as quoted above, that this acquaintance "constitutes a third category of knowledge" that's not reducible to facts or know-how. But labelling knowledge by acquaintance as a third category that's opposed to factual knowledge or skill-based knowledge isn't quite right, and I don't believe, given what he says elsewhere in his "Phenomenal Knowledge," that this is exactly what he means. The problem is this: it's intuitively clear what is meant by 'factual knowledge' or 'know-how', the first being information and the second being abilities. These two things *are* knowledge. However, 'knowledge by acquaintance' is a means of acquiring knowledge rather than being knowledge itself. It is better contrasted with the acquisition of knowledge verbally or through physical practice. It's important to be clear about this, because if it isn't the phenomenal knowledge itself that falls into this third category—if the categories we're talking about are methods of acquiring knowledge—then we're left with a pertinent question: what kind of

knowledge is phenomenal knowledge? Another way of phrasing it: what is phenomenal knowledge knowledge of?

Horgan believes that phenomenal knowledge is still knowledge of physical information. He distinguishes between two types of physical information we may possess. The first kind is what I've been calling 'physical facts', and which Horgan would dub 'explicitly physical information' (1984: 150). Facts of this sort grant knowledge of states of affairs, the kind of stuff that we expect a theory of physics to express. But there's another kind of physical information which Horgan calls 'ontologically physical information', and this includes terms that may have physical referents without the statements themselves being explicitly physical. By 'explicit', I take it that he means something like 'obvious', such that the terms that physics employs are obviously physical, whereas there may be more ontologically physical stuff out in the world that the terms of physics fail to pick out. Horgan claims that Jackson equivocates between these two types of physical information, and I'm in agreement. Jackson lets Mary learn all of the explicitly physical information, which is precisely the type of information one can learn from textbooks, but it doesn't follow from the fact that she learned something new upon seeing a ripe tomato that she learned something nonphysical—she just learned, Horgan claims, ontologically physical information about what this property *is like* (1984: 150-1). Horgan's suggestion is that qualia *may be* physical and not learnable through explicitly physical information. It simply does not follow from the possession of all of the explicitly physical information that one will acquire the ontologically physical information. This is consistent with what Conee is arguing: we must come to acquire phenomenal information through

acquaintance only, explicitly physical information will not suffice. But I believe that these considerations do not establish the ontological status of phenomenal knowledge: we still do not know what phenomenal knowledge is about.

I believe that Conee is right that acquaintance is how we come by phenomenal knowledge. And showing this is powerful: if we can come to this type of knowledge only by acquaintance, then Jackson's knowledge argument fails to prove physicalism false. As he says, "according to this account, Mary already knew all of the physical facts without knowing what it is like to see something red. Mary came to have the latter knowledge simply by having the right sort of experience, and not by acquiring any new information" (Conee, 1994: 141). The knowledge argument, then, doesn't prove that there are nonphysical facts: of course Mary didn't know what red looked like, she simply wasn't acquainted with it. However, what Conee and Horgan say does not show that Mary has learned something physical. I believe Horgan's distinction between the two types of physical information is useful, but it doesn't settle what phenomenal information is. The most we have is the possibility that qualia may be ontologically physical, but nothing we have seen yet has established that. Note that Conee and Horgan are well aware of this. Horgan, in the particular paper referenced here, is concerned only with showing that Jackson has made a mistake, and Conee is stating only that phenomenal knowledge does not necessarily lie outside the realm of physicalism—his project is not to settle the ontological status of the phenomenal (1984: 147; 1994: 140). But it is integral to the current project that we get a grip on this elusive type of knowledge.

The explanatory gap remains. Conee's proposal is a useful way of putting a point that has been recurrent thus far: in order for us to have an experience, or become acquainted with a phenomenal property, we must undergo the right neurological process, as presumably that's the only way that we manage to experience such properties directly. The learning of skills or facts is insufficient, we must undergo "a maximally direct cognitive relation to the experience" (Conee, 1994: 136). That we must have this relationship to qualitative experiences is telling, but I don't believe the relationship itself constitutes this brand of knowledge, for it does not tell us where in the world this knowledge resides. We're left wondering precisely where it belongs in our ontology, and, epistemically, it seems to float above the physical with no obvious attachment to it. I believe that the directness of the relationship between ourselves and what we experience is important in explaining what phenomenal knowledge is, and I shall return to this point in the next chapter.

We have established that phenomenal knowledge is not something we can acquire factually, and so the knowledge argument fails to establish the falsity of physicalism. But all we have managed to do is buy physicalism some time to figure out where to shove the phenomenal character of the world. That we must meet an experience directly to know it tells us nothing of how it fits into the physicalist picture. In order to close the gap, we'll still need to see how the physical world determines the phenomenal, even if it cannot be done through physical facts. Finally, and importantly, we need to know what phenomenal knowledge is knowledge of. This is a question I will provide an answer to in Chapter 2.

Section G: Transition to General Theories

Jackson's knowledge argument is meant to prove that there are facts that are not physical facts. Physicalism, however, should have it that all facts are physical facts. The existence of these non-physical facts, then, should undermine physicalism. In response, others attempt to defend physicalism by denying that phenomenal knowledge is about anything non-physical. As we saw, Dennett attempted to deny the intuition behind the thought experiment, claiming that one really could deduce or otherwise come to know the phenomenal information solely through the possession of the physical information. Lewis took a different route, arguing that phenomenal knowledge is nothing more than know-how, no more mysterious than the knowledge involved in knowing how to ride a bike.

We've seen that these approaches don't work. Dennett, in his attempt at granting RoboMary phenomenal knowledge while she was locked away ended up accidentally conceding the point to Jackson. Lewis' proposal did not fare much better, as we saw that the skills Lewis wished to identify with phenomenal knowledge were neither necessary nor sufficient for the possession of such knowledge.

Conee took a different route. Jackson attempts to establish that phenomenal facts are non-physical facts because they cannot be deduced from descriptions of physical facts. Conee claims that phenomenal knowledge must be learned by acquaintance, not through the possession of facts. Indeed, phenomena are not facts at all, hence why they cannot be learned through lessons and textbooks. In order to become acquainted with phenomena, then, one must have a direct cognitive relationship with them. This is no different than how we come to

know cities. No number of facts could ever acquaint us with Detroit; it is not until we actually experience Detroit that we can be said to know Detroit. Since Mary was never acquainted with redness, no number of facts would have granted her that acquaintance. Nothing short of seeing red itself would work. Thus, the knowledge argument fails. It isn't that phenomenal knowledge must lie outside of the physical world, but rather it simply cannot be acquired through facts.

We have not, however, saved physicalism. At best, we have bought it some time. That phenomenal knowledge cannot be acquired through facts does not say anything about whether it lies within or without the physical world. The knowledge argument may not have been successful, but physicalism is still at risk. There is a crucial question at hand: if phenomenal knowledge is not knowledge of the physical world, then what is it knowledge of? The answer to this question, I submit, depends on where the phenomenal belongs in the physical world, if indeed it does. Thus, we need an account of where in the world we can place the phenomenal. Philosophers of mind have been attempting to answer this very question, and the major options that have been offered have been dualism, behaviorism, the identity theory, and functionalism. Dualism is out, since it is not a physicalist theory. Thus, we are left with the latter three.

The objective now, then, is to consider briefly these three theories. We must see whether they offer a plausible account of where the mental belongs in the world—of what the phenomenal is supposed to be. I argue that they do not succeed. These theories all fail to give an adequate account of the phenomenal's place in the world. In other words, none will succeed

at closing the explanatory gap, and no amount of modification will ever suffice for doing so. I will argue that they all fail for the same reasons, and that these reasons are instructive. Using the lessons we learn from the pitfalls of the theories we'll consider, I'll make clear what a theory of mind would need to do in order to succeed at finding a place in the world for mentality. Making a place for the phenomenal in the world—that is, providing a physicalist answer to the question "what are phenomenal properties?"—will allow us to answer what they are knowledge of. With these questions answered, I will be well-situated to take on the task of closing the explanatory gap.

Section H: Behaviorism

The objective of behaviorism—specifically logical behaviorism—is to explain mentality in terms of behavior in the hopes of bringing the field of psychology within the scope of physics. The motivation stems largely from the perceived chasm between the subject matter of psychology and that of the physical sciences. Physics is understood as dealing with the external, objective world. It deals with intersubjective phenomena, and its statements are subject to empirical verification. Not so for psychology, supposedly. Psychology was seen as dealing with the inherently meaningful and purely subjective, and so it was seen as lying forever beyond the reach of empirical observation. Carl Hempel and other logical behaviorists believed that psychology, like the other hard sciences, could have its statements made available to empirical verification, thereby closing the gap between the two.[8]

[8] Later in his life, Carl Hempel no longer held the views he expressed in the paper I cite.

Before delving into how we should conceive of psychology, Hempel aims to "clarify the very concept of the subject matter of science." According to Hempel, "the theoretical content of a science is to be found in statements. It is necessary, therefore, to determine whether there is a fundamental difference between the statements of psychology and those of physics" (1980: 17). The idea is that physics is in the business of making a certain kind of statement, namely, statements that are open to verification. Logical behaviorism is, after all, a verificationist theory, and it adheres to the verificationist doctrine: "The meaning of a statement is established by the conditions of its verification" (Hempel, 1980: 17). In other words, what makes the statements of physics meaningful is that they give rise to certain test conditions that can be verified. What it means for any particular statement of physics to be true, then, is that the test conditions that it expresses all prove the statement to be true.

In order for the logical behaviorist to be successful, the statements of psychology must be of the same kind as the statements of physics. Thus, the task is to translate the sentences of psychology, which contain subjective, mental terminology, into verifiable sentences that lack that terminology without loss of meaning. These new sentences must provide the test conditions for their verification. Hempel believes we can do just that. Consider, first, what this would mean in the field of physics. Hempel utilizes a thermometer as an example, and so I shall do the same. Take the statement "it is 70 degrees Fahrenheit in this room." This statement, claims Hempel, could be translated into a series of what he called 'test sentences' that express the relevant test conditions. These sentences would look something like this: (1) The functioning thermometer on the wall reads "70 degrees Fahrenheit"; (2) A different device

within the same room reads “21.11 degrees Celsius”; (3) Materials of such-and-such composition have expanded to such-and-such state; etc. The list would presumably be very long, but given that the test sentences all turn out to be true, then the original statement will be true as well. Indeed, the original statement is nothing more than a simplified manner of expressing this far longer conjunction of test sentences.

We can do the same with psychological statements, Hempel believes. As an example, he asks us to consider a man named Paul who has a toothache. The statement “Paul has a toothache” contains a mental term: ‘toothache’. But, this statement is also just a simplified expression of a much longer conjunction of test sentences that can all be verified. This conjunction would look something like this: (1) Paul winces and moans; (2) Paul has certain physiological states and happenings; (3) Paul expresses statements such as “I have a toothache”; (4) Paul has a decayed tooth; etc. These statements do not contain any mental terminology, and together they form the test conditions for the original statement’s verification. Thus, “the statement in question, which is about someone’s ‘pain’, is therefore, just like that concerning the temperature, simply an abbreviated expression of the fact that all its test sentences are verified” (Hempel, 1980: 18). We can do this for all of psychology’s statements that include mental terminology, argues Hempel, and so there is no in principle difference between the statements of physics and those of psychology. Something worth noting: the claim is not that all there is to mentality is these particular behaviors, but rather that these translations *mean* the same thing as their original statements.

These translations won't work. In the case of Paul's toothache, there may be some set of sentences that we can provide that make it appear plausible that a complete translation is within reach. However, in order for logical behaviorism to be successful, it must be the case that *all* sentences containing mental phenomena can be translated in this way, and there are significantly more complex mental states that we experience beyond toothaches. Someone might believe Goldbach's conjecture to be true, and it's completely opaque which set of behaviors would serve as the appropriate test sentences for this belief. Similarly, one could believe all manner of things about mathematics, relativity, the interactions between black holes and neutron stars, and so on. Translation of such mental states seems hopeless, though we may be tempted to appeal to verbal reports as a starting point. However, it turns out that the logical behaviorists cannot avail themselves of this type of report.

Consider again the translation of the sentence concerning Paul's toothache. The objective is to translate that sentence into a set of test sentences that lack mental terminology. This is absolutely of the essence, as Dennett articulates: "no satisfactory psychological theory can *rest* on any use of intentional idioms, for their use presupposes rationality, which is the very thing psychology is supposed to explain. If there is progress in psychology, it will inevitably be, as Skinner suggests, in the direction of eliminating ultimate appeals to beliefs, desires, and other intentional items from our explanation" (1978: 68). But sentences such as "Paul expresses sentences such as 'I have a toothache'" serve only to hide the mental phenomena upon which they rest. Indeed, we can take Paul's statement of "I have a toothache" as a verifiable test sentence only if we *presuppose* that Paul *believes* himself to have a toothache and

desires to be truthful. Otherwise, the statement is useless as a test sentence. The same will hold true for *all* verbal reports, rendering them all useless to the logical behaviorist. Without verbal reports, translating complex mental states becomes a genuinely hopeless task.

Now, Hilary Putnam argues that the complete translation of sentences with mental terminology into sentences without mental terminology is impossible, though few logical behaviorists held the view to such an extreme for very long. He offers a milder, more acceptable version of the claim:

> ...there exist entailments between mind-statements and behavior-statements; entailments that are not, perhaps, analytic in the way in which 'All bachelors are unmarried' is analytic, but that nevertheless follow (in some sense) from the meanings of mind words [....] These entailments may not provide an actual *translation* of 'mind talk' into 'behavior talk' [...], but that this is true for such superficial reasons as the greater ambiguity of mind talk, as compared with the relatively greater specificity of overt behavior talk (Putnam, 1963: 25-6).

This version of logical behaviorism is more difficult to object to than the variety defended by Hempel, though we'll see that it, too, cannot work.

According to Putnam, the project of the logical behaviorist is impossible. Part of the reason is that the behaviorist is aiming to make mental facts out to be logical constructions out of behaviors. However, mental phenomena are the *causes* of behaviors, and causes cannot be logical constructions made out of their effects. The sensation of pain, says Putnam, is the reason that one behaves as though one is in pain. He believes that "pains are not clusters of responses, but that they are (normally, in our experience to date) the causes of certain clusters of responses" (Putnam, 1963: 28). It would be a mistake, however, to believe that just because

these two things are normally correlated that there's any relationship of necessity between them. To drive the point home, he asks us to consider a thought experiment. I will present a very condensed version below.

Imagine there's a race of beings called 'Super-Spartans'. These Super-Spartans are very similar to us, but they never exhibit pain behaviors. They do feel pain, they just find it deeply shameful to reveal any kind of pain response. To avoid the logical behaviorist appealing to the behaviors of the children of the Super-Spartans, he envisions that they have evolved such that even the offspring is completely acculturated, so that none of the members of the species ever reveal pain behaviors, even when they are in the most excruciating pain. At this point, the logical behaviorist may be tempted to simply say that they are actually not in pain. But Putnam asks us to consider a scenario in which we meet one of them, and that this particular Super-Spartan adapts to our culture and ends up admitting that he used to feel intense pains just as much as he does now, but now he is free to express himself. It would be absurd, claims Putnam, to believe that were it not for this singular Super-Spartan, the rest of the Super-Spartans would be pain-free.

Putnam's thought experiment becomes more outlandish with each iteration, appealing to bizarre X-waves and other waves emanated by the Super-Spartans in an attempt to deceive us. To simplify: Putnam's point is that the behaviors of the Super-Spartans may tell us very little about their mental lives. There could be all manner of ways they may behave that may or may not correspond to our own behaviors, and it wouldn't be clear what they are feeling. This is a point that will come up again with the relevant adjustments: the behaviorist's theory does

not tell us anything about pain, as other beings who are less like us may not behave as we do, and it would be impossible to infer what they're feeling.

This lack of an entailment relation between mentality and behavior, even a weak one, exposes a gap in explanation. Indeed, even if we could successfully get a list of behaviors that would guarantee the presence of certain mental states in a given species—that is to say, even if there existed a species whose behaviors just so happened to perfectly align with certain mental states—we would still face the gap. This would fall short of the logical entailment relation Putnam rightly claims is required, for there would still be nothing about the behaviors of that species that would necessitate any given mental state. Indeed, we wouldn't even know that the behaviors and mental states of the members of that species lined up so perfectly; intuitively, all of the wincing and moaning is logically consistent with any sensation at all. Compare this to a different, successful case of the requisite entailment: "Take, for example, the case of the man who speaks. Within the framework of physics, this process is considered to be completely explained once the movements which make up the utterance have been traced to their causes, that is to say, to certain physiological processes in the organism, and, in particular, in the central nervous system." In this particular case, there is clearly nothing left to explain. To have spoken *just is* to have gone through the relevant physical process. Even if we turn out to be fundamentally wrong about physics, our current physical models would still provide the same picture, and there would be no question that things simply could not be any different. We cannot hold the physical fixed and vary the act of speaking. But the same would not hold for mentality if logical behaviorism were true. Holding behaviors fixed, not only could mental

states vary between worlds, but it seems clear that they can and probably do vary between individuals in this world.

So logical behaviorism cannot close the gap. That may not be particularly surprising, but the reason is informative: phenomenal states and the proposed explanatory behavioral states do not bear the appropriate relationship. It needs to be the case that the proposed explanation—in this case behaviors—leaves nothing dangling loosely. An appropriate explanation should rigidly tell us which phenomenon to expect in the same way that the physiological explanation tells us precisely what occurred. In more familiar terms: phenomenal knowledge is not knowledge of behaviors; the act of wincing and groaning is not what pain *is*. Logical behaviorism fails to provide a place for phenomena in the world.

Section I: The Identity Theory

We have already talked a bit about the identity theory at the beginning of this chapter. To reiterate, the identity theory is the view that mental properties are identical with physical properties. In what follows, I will speak largely of mental and physical states for ease of exposition, though talk of states, processes, and events tends to be cashed out in terms of property identity. That said, the identity theory has it that every mental state—pain, pleasure, etc.—is identical with some neural state in our brains. In discussing the identity theory, we looked at a powerful problem first articulated by Kripke. The problem ran like so. The terms 'pain' and 'C-fibers' in the statement "pain is C-fibers firing" are rigid designators. Identity statements between rigid designators must be necessarily true if they're to be true at all, for the

terms pick out the same entities across all possible worlds. Yet, it seems obvious that pain is not *necessarily* C-fibers firing. So, runs the *modus tollens*, pain is not the firing of C-fibers in this world either. As we saw when we first ran through this, the identity theorist has a way out: bite the bullet.

While it is possible for the identity theorist to claim that pain and C-fibers firing really are necessarily identical, regardless of the counterintuitiveness of the claim, there are much larger problems that the identity theory has to confront. The first we have already encountered before, and that is the problem raised by Levine: the theory tells us nothing about why a given physical state should feel any particular way. The problem here is largely predictive. Even given a full listing of every mental-physical correlation in humans, we still wouldn't know how much the physical states can vary while still counting as instances of the same mental state. For example, neural state A may be pain, but nearly identical neural state A* can remain a complete mystery. It might, phenomenally, be almost exactly the same as pain or entirely different. The identity theory simply doesn't tell us. If this seems like a familiar problem, it's because it's a different articulation of the same problem that plagued behaviorism above: no amount of physical information appears to be capable of settling phenomenal information. We find ourselves dealing with the explanatory gap once again.

I take it that this alone is enough to show that the solution to our problem must lie elsewhere, but Putnam raises a stronger objection: the identity theory is incompatible with instances of multiple realizability. Now, William Bechtel and Jennifer Mundale believe they have an adequate response to Putnam's objection that not only solves the problem he raises,

but at the same time grants the identity theory predictive power. If they're right, then the explanatory gap is closed, for physical information actually would shine a light on phenomenology. We'll take a look at their argument, and then I'll respond that they not only fail to quash Putnam's concerns adequately, they seem to concede the point inadvertently. First, however, I must briefly construct the problem of multiple realizability.

From the outset, it seems obvious that human beings are not alone in possessing certain mental states. Indeed, we believe that pain is likely felt by non-human entities: dogs, bats, octopuses, and maybe even aliens and robots. Now, if we take the identity theory at its word, then pain is literally the firing of C-fibers. But it's quite unlikely that octopuses have C-fibers. Perhaps they have D-fibers. If the identity theory is right, then the result is that C-fibers are D-fibers. But this is absurd, and so the identity theory is in hot water. The idea is simply that mental states are multiply realizable: they can have more than one physical realizer. Human neurons, octopus neurons, and maybe even hyper-advanced circuitry can all give rise to the same pain. But if we draw an identity between the pain and its realizers, then human neurons, octopus neurons, and circuits are all identical. Standard solutions appeal either to disjunctive identities or the promise of some more specific neural (or otherwise) structure shared by all creatures. These are live options, though the latter seems unlikely, and the former leaves us with an explanatory gap.

Bechtel and Mundale begin their argument by stating that it is precisely because of the close similarities between different creatures that we can draw inferences about their mental states. They state that it is, in fact, "the very *similarity* (or more precisely, *homology*) of brain

structures which permits us to generalize across certain species. So [...mental states] are not multiply realized" (Bechtel and Mundale, 1999: 178). Their claim is basically that the identity really does hold, and we can see this in the similarities that exist between different creatures in similar mental states. So, it is not the case that mental states can be multiply realized; all of the evidence we have points to the same structures giving rise to the same mental events. If this seems wrong, it's only because we tend not to be careful with how thickly or thinly we slice our concepts. Philosophers, they claim, have a tendency of slicing mental events very thickly and physical/neural events very thinly. This gives rise to the misconception that mental states are multiply realizable, since we are asking a very thin brain state to account for something as thick as 'pain' across creatures. In order for Bechtel and Mundale to be successful, they must show two things: first, when we fix the grain of our concepts, we should no longer run into problems of multiple realizability; second, in order to save the theory from the explanatory gap, it must be capable of making predictions. They believe they achieve both, but there are some serious problems ahead.

Before we get into some of the more worrying objections, a quick note on identity. Bechtel and Mundale are too loose with their treatment of identity. They state that "neuroscientists will attempt to identify the same brain areas and same brain processing in different members of the same species and across species *despite whatever differences there are*" (Bechtel and Mundal, 1999: 201-2, *emphasis mine*). This is problematic if we aren't very careful. The identity claim of the identity theory does not allow for any differences. Whatever it is that we ultimately identify pain with—be it a very complex structure, a less complex one, and

be it made of neurons or sludge—that thing must be present exactly as it is across all tokens of pain. Otherwise, we have broken the identity. The task of identifying may be difficult, but the point is just that regardless of what we ultimately learn pain to truly be, that thing must exist everywhere there is pain of the same type. The way Bechtel and Mundale express themselves in the above quotation is more reminiscent of functionalism than it is of the identity theory. If what we are concerned with is identity, then we must be very strict.

Bechtel and Mundale set out to quash multiple realizability concerns because if they hold water, then that means that the identity theory is in trouble. However, the response they provide does not absolve the identity theory. Indeed, depending on how we spell out the details, either the identity theory still has to deal with issues of multiple realizability, or it is instead subjected to a different problem.

Remember that the existence of so much as a single shared mental state with more than one physical realizer would be unacceptable. As stated previously, Bechtel and Mundale insist that multiple realizability is really just a problem of grain size. Specifically, the issue arises only because we slice the mental far more thickly than we do the physical.

> When comparing psychological states across different individuals, psychologists also tend to ignore differences and focus on commonalities. Likewise philosophers such as Putnam, who proposed comparing psychological states such as hunger across species as remote as humans and octopi, have abstracted away from differences. Yet, at anything less than a very abstract level, hunger is different in octopi than in humans. So, just as neuroscientists abstract away from differences between brains in identifying brain areas and brain processes, so do psychologists and philosophers in identifying psychological states (Bechtel and Mundale, 1999: 202).

The idea, I take it, is this: Even within a single human being, pain comes in many varieties. We have pinpricks, sore throats, headaches, the pain of a ruptured appendix, and so much more, and this is to say nothing of emotional pains. These are all under the very thick umbrella of 'pain'. All of these pains, within the same singular organism, likely have different physical realizers, different brain states that bring them about. Yet we wouldn't take this as an objection to the identity theory. Well, the same is likely to hold true for the hunger an octopus feels as compared to our own. If we just slice the mental state as thinly as we do the physical, we'll see that we're actually dealing with two different mental states *and* two different physical states. We can lump them both together under the same thick concept of 'hunger', but then we'll need to ensure we carve up the physical types just as thickly. Once we do this, we should see that we no longer have to worry about multiple realizability.

The only way that we can carve things up evenly would be to carve the mental as thinly as the physical and not the other way around. This is because in order to carve the physical thickly we would need to abstract from the particular physical facts, and the identity theory cannot allow for this. This, however, would result in an explosion of phenomenology. Now, it is far from problematic to believe that octopuses and humans probably feel hunger differently, but the claim that the mental state must be carved just as thinly as the physical state results in two problems. The first is just that we must carve physical states *very* thinly, and for each one of those slices, there must be an equivalently thin mental state. If there isn't, we'll run into multiple realizability again. This means that for the displacement of a single neuron, even a single electron, there must be a corresponding mental state. Now, Bechtel and Mundale do not

specify just how thin the physical states would have to be, so perhaps we can opt for a thicker slice that could allow for subatomic variation. This, however, cannot work within an identity theorist framework. The identities that the type identity theorist draws are not merely structural; if they were, then the theory would be functionalist. Consider: we draw the identity at the level of neurons and ignore their components. If we can ignore their components, then we can have two structurally identical neurons with different material components (perhaps the H_2O molecules within have been replaced with XYZ). What would hold the same between the neurons would be structural and functional facts, but strictly speaking, there are two physical states here. If they both give rise to pain, then we have multiple realizability again. Identity theorists cannot avail themselves of abstract identifications.[9] The identity is a physical one, meaning that the mental state must be identified with the physical state down to the last quark. The identity needn't include *every* particle in the brain, only those relevant to the brain being in the state that it is in. However, any one neuron that gets included in the identity must have its entire constitution included. If just one proton is relevant, or one molecule of thymine, then it must be included in the identification. So, identity theorists cannot avail themselves of particularly thick physical states, and variation in so much as the position of that molecule of thymine must result in a different mental state.

Perhaps this is less problematic if we take the mental differences to be extraordinarily minute. In a moment, we'll see precisely why we cannot do that. But consider the second

[9] At least not at the macrophysical level. One can embrace a microphysical identity theory, though most identity theorists wouldn't and Bechtel and Mundale don't have such a theory in mind. The theory I put forth in Chapter 2 is such a theory.

problem that arises. This explosion of phenomenology actually threatens to bring the problem of multiple realizability with it. The more physical/neural states there are, the more mental states there must be, but the more mental states there are, the more pressure there is to believe that some of them can have more than one type of physical realizer. If there are 400 billion trillion types of pain that neurons can bring about, it's not clear why we wouldn't be able to recreate so much as a single one in a sufficiently complex computer chip. These concerns alone do not show that Bechtel and Mundale's solution cannot work, but there's more trouble ahead.

The solution that Bechtel and Mundale propose does not provide the predictive power they claim it does. Indeed, they seem to acknowledge this themselves. As stated above, they believe that humans and octopuses share the experience of hunger only in the most abstract sense. I take this to mean that the qualitative feel of the hunger itself is very different between the two. Indeed, the motivation even to call this 'hunger' is likely just a behavioristic tendency. We see how they behave around food, and so we call it hunger. But Bechtel and Mundale have absolutely no basis for this claim. Note, the claim is not that these hungers may be somewhat different sensations. The claim is that the term 'hunger' can apply only to both in a very abstract sense, after we have removed all of the differences between the two. This is a surprising claim, since they spend considerable time talking about the contribution that neuroscience can make to psychology. Philosophers, they claim, have long argued that the pesky physical facts of neuroscience have nothing to offer psychology, since the phenomenal is multiply realizable. This is just the issue Levine mentions and that we've already seen: the physical does not settle

the mental. Well, if the physical cannot settle the mental, then the physical cannot offer predictions about how physical things are phenomenally. The objective, then, was to bring the mental back within the realm of the physical. The point was to show that neuroscience *can* make predictions on these shared physical structures in the brain. But to say that the octopus and human sensations of hunger share almost nothing in common is to give the game away. It is to say, once again, that neuroscience has nothing to offer psychology.

Perhaps identifying these similar brain structures is a great way of predicting behavior or picking out a function. But to say that the sensation of hunger just *is* the neural state grants us absolutely no information. We started by thinking that the octopus and the human are both hungry. We then look into their brains. Admittedly, there are differences between the two brain states. So, we abstract away from those to find what holds the same. What we're left with are actually two different brain states. Once we slice the mental states just as thinly, it becomes clear that we have no idea what the octopus is feeling. It might be almost identical to the human sensation of hunger, or it might be something utterly alien. We are once again left in the dark. I think this is where Bechtel and Mundale accidentally concede that, at best, neuroscience can provide us with behavioristic or maybe even functional explanations, but it cannot help the identity theory. Notice that this worsens a problem I mentioned previously. Perhaps, I said, an explosion of phenomenology would be less problematic if minute differences in physical states resulted in minute differences in mental states. But we have no reason to believe this to be true. The neuroscientists cannot even tell us how much variation in mentality we can expect from small changes in the physical, and so the displacement of a single electron could result in any

sensation at all. Note, I don't think that we should believe this nor that it's a consequence of the identity theory that this happen; rather, the point is just that this possibility remains open, and the identity theory can give no reason against it.

And so we're left in the dark with a theory that does not bridge the gap between the physical and the mental. Indeed, it appears that we may be the worse for wear after we consider the consequences of the arguments presented above. The structural similarities that were supposed to be illuminating have left us in a world with myriad mental phenomena and no cohesion between them and their realizers. It's not clear why on the basis of neurochemistry we should believe the octopus to be feeling anything at all.

Much of what Bechtel and Mundale have to say would be better suited, I believe, to a defense of functionalism. Functionalism is a far more promising theory that seems capable of explaining mental phenomena more completely. We'll end the chapter by considering what it has to offer, and ultimately arguing that it, too, cannot cope with the problem raised by the explanatory gap.

Section J: Functionalism

Functionalism is the view that every mental state is identical with some functional state. More precisely, it's the view that mental states act as the causal intermediaries in a functional system between inputs and outputs. One of functionalism's strengths is its ability to include other mental states in its functional explanations. For instance, given the input of tissue damage, the functionalist can allow for behavioral outputs and for mental changes, such as anger. This is a

powerful improvement over behaviorism, which cannot allow for the use of mental terminology in its theorizing. Additionally, because functionalism identifies mental states with functional relations between inputs and outputs in a system, it avoids the problems of identity that the identity theory faces when dealing with multiple realizability.

There are quite a number of different species of functionalism, though they can be broadly categorized into two major groups. Following Ned Block, we can call these two groups 'Functionalism' and 'Psychofunctionalism' (1978).[10] Functionalism is solely concerned with characterizing mental states by their relations to observable inputs and outputs, where the specification of those inputs and outputs is "plausibly part of common-sense knowledge" (Block, 1978: 269). The inputs and outputs themselves are expected to be outside the body, where the inputs are external things that occur to the system, and the outputs are behaviors of the body itself. Psychofunctionalism, on the other hand, can allow for internal occurrences, such as neural firings, to count as its inputs and outputs.

Lewis provides perhaps the most well-known version of Functionalism. According to Lewis, mental states are understood as the causal mediaries between inputs and outputs specified by folk psychology. For instance, we all have a pretty good understanding of what types of inputs and outputs to expect when someone is in pain. As an input, one may have jammed one's toe on a coffee table, and the expected output might be a yelp. We must take the entirety of our folk psychological understanding of mental terms, and using the commonly understood relations between inputs and outputs, we can construct a functional theory of

[10] Ned Block uses the capitalized terms to differentiate between the two and uses the lowercase term 'functionalism' when speaking about the theory more broadly. I shall do the same in this section.

things like pain. Once we identify what it is in the world (or, more precisely, our brains) that perfectly fits the role of pain in our psychological theory, we will have successfully found the realizer of pain or the pain itself, depending on the functionalist theory. To elucidate, Lewis presents an analogy about a detective constructing a theory about a crime. The detective, using variables to name the potential culprits, details an elaborate story about the events the culprits were involved in. They caused such-and-such happenings and had such-and-such happenings occur to them. The detective, in telling the story, "set forth three roles and said that they were occupied by X, Y, and Z [...] They were introduced by an implicit functional definition, being reserved to name the occupants of three roles. When we find out who are the occupants of the three roles, we find out who are X, Y, and Z" (Lewis, 1972: 251). In other words, if there turn out to be three unique individuals who did do all of the things we theorized about X, Y, and Z, then we have found who X, Y, and Z are. Similarly, we theorize about mental states by claiming that they are the things that are caused by such-and-such and result in such-and-such. If we find some particular state of the brain, or the body, or what have you that successfully manages to map the inputs to the outputs as detailed in our theory, then we will have found the realizer of that mental state. Notice that this can be to whatever level of abstraction is necessary; so computer chips may be just as apt as brain matter, so the theory claims. The way theory construction is supposed to work is technical and requires the replacement of mental terminology with variables. The complications are not necessary for my purposes. What matters is that "when we learn what sorts of states occupy those causal roles definitive of the mental states, we will learn what states the mental states are..." (Lewis, 1972: 256).

Psychofunctionalism does not touch upon folk psychology. Instead, what matters is largely what the field of psychology tells us. A mental state, then, is "the state of receiving sensory inputs which play a certain role in the Functional Organization of the organism. This role is characterized, at least partially, by the fact that the sense organs responsible for the inputs in question are organs whose function is to detect damage to the body..." or whatever else, and provides outputs of, for instance, aversion due to the "high disvalue" the organism assigns to the inputs (Putnam, 1990: 55). This grants Psychofunctionalism greater freedom in what it treats as relevant, taking into account not only inputs and outputs, but also "causal relations among whatever psychological events, states, processes, and other entities [...] actually obtain in us" (Block, 1978: 274).

The primary motivation behind functionalism was the restrictiveness of the identity theory. Because the identity theory tried to identify mental states with human neural states, it had as a consequence that other beings that did not possess those precise neural states could not count as having the appropriate mental state. In Block's terms, the theory was too chauvinistic. It's deeply implausible that there must be some one particular neural state that is shared among all things that can feel pain. The functionalist, then, wished to provide a theory that could allow other things to possess mentality as well. In other words, functionalists wanted a greater degree of liberalism. Whereas a chauvinistic theory denies mind-possessing entities minds, a more liberal theory should grant them those minds. Functionalism does not wish to be too liberal, however. Being too liberal would result in granting entities that do not possess minds mental states. Block argues that functionalism, rather than escaping chauvinism,

will either end up being too chauvinistic or too liberal depending on how it's fleshed out. Indeed, there will be no characterization of functionalism that avoids both chauvinism and liberalism.

We will now take a brief look at the concerns that Block raises for functionalism. I will be largely focused on the intuitive thought experiments Block appeals to in articulating his objections, as I believe them to be quite powerful. Block himself does not believe his objections to be decisive, but I think his thought experiments do reveal something important about why functionalism will fail to close the gap: it does not give adequate attention to the importance of qualia.

Let's begin by considering why functionalism may be too liberal. According to Functionalism (with a capital 'F'), what matters to the ascription of mentality is that the inputs and outputs of the system correspond to our folk-psychological understanding of mental terms. So long as a system takes in the right inputs and dishes out the right outputs, the system counts as having a mind. Let's now imagine that we manage to get one hundred billion people together to participate in an interesting experiment.[11] We'll give each of them a walkie talkie, and we'll have them simulate the exact functions of some particular human brain. We will then hook up the inputs and outputs of this system to a human body, such that the body will behave in all of the relevant ways. So, if the body stubs its toe, that information will be sent to the relevant people who will then, in turn, send their signals via walkie talkie to other people, until the entire system provides output back to the body, which will in turn yelp. As Block

[11] Block originally appeals to the Chinese nation, but our brains, as we now know, have about one hundred billion neurons, so I'll be using this modified version.

says, “surely such a system is not physically impossible,” and “it could be functionally equivalent” to a human being for some period of time (1978: 279). Yet, clearly, such a system does not have a mind. The problem with this type of functionalism is that it permits too many things to count as having minds. Block constructs another thought experiment in which a rich sheikh takes control of some economic system and makes it behave in a functionally identical manner to a human mind (1978: 314-5). Economic systems do, after all, have inputs and outputs, and it is far from inconceivable that this could be achieved, even if it is wildly impractical. So long as the system behaves in the appropriate manner, Functionalism must ascribe it a mind.

Functionalism needn’t be so abstract that it allow for economic systems to count as having minds. It can be constructed in a more restricted manner. However, then it will fall victim to chauvinism. Indeed, Block believes most versions of functionalism turn out to be guilty of chauvinism rather than liberalism: “Functionalists tend to specify inputs and outputs in the manner of behaviorists [...]. Such descriptions are blatantly *species-specific*. Humans have arms and legs, but snakes do not—and whether or not snakes have mentality, one can easily imagine snakelike creatures that do” (1978: 316). Any form of functionalism that ends up requiring that things be like us will inevitably leave certain creatures out. Both Functionalism and Psychofunctionalism will be chauvinistic if not liberal. The Functionalist that appeals to yelp-like outputs will inevitably fail to count those creatures that cannot yelp as being in pain. This would include the Super-Spartans that Putnam used to object to behaviorism. The neural outputs of Psychofunctionalism will inevitably leave out those without neurons. Abstracting

won't help, because it will either still leave out creatures less similar to us, or else it will reach a level of abstraction that will include too much.

Block rightly believes that the battle to construct a form of functionalism that manages to capture all and only minds is futile. Any being we manage to create a perfect functional description of will always leave out easily imaginable cases of other beings who are just different enough for the description to wrongly leave out. Indeed, it seems that we are capable of imagining all sorts of mind-possessing beings that completely fail to resemble any offered functional description. It appears to us that almost any set of inputs, outputs, and processes could belong to a mind-possessing entity.

While Block does not believe his objections to be decisive, there's something important that his considerations and thought experiments bring out. In constructing his thought experiments, he is using an intuitive notion of what gets to possess a mind. Even when constructing formal arguments, the concern is always that some things will have minds that don't fit within the functional description, or some things will lack minds that do fit the description. But what drives this intuition? I submit that it is the presence of qualia. What we are imagining when the sheikh commands the economic system is that the system has no qualitative experience. Whatever else may be true of the system (e.g., perhaps, under certain definitions, the system can *believe* things), it cannot experience anything at all. Similarly, we are envisioning that functional descriptions can fail to capture radically different entities that *do* have qualitative experience. Regardless of whether our intuitions about which systems do and don't have qualia are right, one thing seems certain: *if* a functional description leaves

something that does have qualia out, it fails, and if it includes something that does not have qualia, it fails. What this suggests is that functionalism, at best, is an attempt to track qualia and not an explanation thereof. But there is more to be said on this point.

Assume that we actually do manage to find the perfect functional description that does capture all and only those things that intuitively have minds. Now, let's consider the possibility of inverted spectra. "It makes sense," says Block, "to suppose that objects we both call green look to me the way objects we both call red look to you. It seems that we could be functionally equivalent even though" our actual color experiences are quite different (1978: 304). Block goes on to imagine lenses that can achieve this, and a pair of twins, one of which was given the lenses at birth, who are functionally identical even though their experiences are different. If this is possible, then one thing is clear: functionalism does not settle phenomenal information. We can hold the functional fixed and vary the phenomenal.

Responses to this possibility tend to be an attempt to point out that there actually would be a functional difference if the qualia were varied. In other words, the imagined scenario is impossible. I find this response unacceptable. One might want to say that if our color experiences were different, then we would have behaved differently from how we do.[12] However, thinking of the inverted spectrum in terms of colors is meant only to be illustrative. That a functional difference would be necessary seems more difficult to believe when we include the other senses, such as sound, smell, and taste. Indeed, senses like taste and smell seem so subjective, it's hard to see how there'd be any functional difference at all. In other

[12] See Shoemaker (1982).

words, it's perfectly conceivable that two functionally identical systems have different gustatory sensations. We could even ensure that the differences be only minor, such that we don't swap any delicious tastes for abhorrent ones. This is not necessary, as whether we find something delicious or abhorrent already does not appear to be tied to the taste itself, but rather some appreciation we have for the taste, but this is tangential.

To defend the possibility of inverted spectra further, let's imagine a scenario. We know that bats echolocate, though we have no idea what that experience is like for the bat. Let us assume that what happens when the bat echolocates is that a mental image is conjured. Perhaps the image is a map of the space and obstacles around the bat. Now, to convey the needed information to the bat, the map need only be in grayscale. Yet, it seems perfectly conceivable that there be one bat whose map is in grayscale and another bat whose map is in monochromatic red. Neither map grants any more or any less information to the bat. We can further imagine both bats to be completely blind, such that they have never seen any other colors (not that they could communicate about their experiences in the first place). The two would be functionally identical, yet there would be a qualitative difference in their experience. This, I take it, is a clear instance of a difference in qualia without a difference in function.

That there could be inverted spectra proves that functional information cannot settle phenomenal information. The functional simply does not elucidate what the phenomenal is supposed to be. So, once again, it appears that physical facts—in this case, functional facts—are incapable of determining the phenomenal. So, the explanatory gap remains.

Section K: Desiderata

It's time to take stock. The theories we have considered in this chapter have all failed to settle phenomenal information for one reason or another. If a physicalist theory cannot settle phenomenal information, then it fails to close the explanatory gap. If it turns out that a physicalist theory cannot close the gap, then the theory is very likely to be false. What, then, would be required to close the gap?

There appear to be four necessary conditions that a theory must meet if it is to have any hope of being successful. First, the theory must be a physicalist theory; second, it must take qualia seriously; third, the theory must provide a reductive explanation of phenomenology; fourth, the theory must require only minimal brutality. Let us take these in turn.

The first two desiderata are straightforward. That the theory must be a physicalist theory is necessary *ex hypothesi*. The theories we have considered thus far have all met this condition, but the theories we will consider moving forward may fail to make sense of why they should be considered physicalist. Also, any theory that is eliminativist will not count as closing the explanatory gap. I take this to be for at least two reasons. The first is that denying one side of the gap is not a solution, it's merely a refusal to confront the problem. The second is that the existence of qualia is undeniable. This dissertation is not directed at those who consider themselves eliminativists about qualia.

Concerning reduction, upper-level phenomena like our experiences must be explained in terms of some type of lower-level, physical phenomena. If a physicalist theory cannot do this, then the phenomenal will remain unexplained. Realize that it may be possible to provide some

other type of explanation for the phenomenal, but such explanations will end up not being physicalist, such as the existence of souls. We don't need a fully fleshed-out reductive explanation. Rather, what's necessary is that it be clear how the reductive project is possible. If the phenomenal cannot be reduced, then it cannot be explained in physical terms, and so the gap shall remain.

Finally, while all theories admit of brute facts, we must be cautious with how we employ them. The idea here is to avoid the type of brutality that makes the identity theory unappealing. Brute facts should be posited only when necessary to provide an explanation and when the fact posited does not itself call out for further explanation. In the case of the identity theory, that pain is the firing of C-fibers was made to be a brute fact. Yet, we remained deeply unsatisfied, for it seems clear that a desire for more explanation is warranted. Similarly, to utilize an example used by Levine, were we to claim that gravitational pull is a brute fact about the world, we would be unsatisfied, for more explanation seems to be clearly required. Yet, when we claim that the gravitational constant, *G*, is a brute fact about the universe, we require no further explanation. Some things must be brute, such as the charge of electrons. We must ensure that any brutality we make use of is similarly warranted and does not beckon for further explanation. The motivation here is clear: the gap we are interested in closing is explanatory in nature, and so if we make use of brute facts that beckon for greater explanation, we will fail.

Even if we succeed in finding a theory that meets these conditions, together they are not sufficient to close the explanatory gap. This is because a theory can satisfy these conditions and still not tell us the exact relation between qualia and the physical. We might, for instance,

make the phenomenal out to be a special kind of property of physical matter, and yet, without saying more, the explanatory gap will remain.[13] There is one more thing that we will need in order to successfully close the gap, and that is an answer to the following question: what is phenomenal knowledge knowledge of? In other words, what exactly are qualia?

Chapter 2 is dedicated to finding a theory that meets the four necessary conditions above. I argue that the only type of theory that shows any promise is panpsychism. Chapter 3 is dedicated to granting qualia a specific place in the world. I will aim to show what phenomenal knowledge is knowledge of. I will defend the view that phenomenal knowledge is knowledge of the intrinsic nature of matter. As is clear, there is much work to be done.

In what comes next, we will be looking at panpsychist theories. My objective is to find a version of panpsychism that meets the four necessary conditions we have articulated. As we will see, most forms of panpsychism will fail to meet our conditions. Ultimately, I will argue that only one type of panpsychism shows promise: micro panpsychism.

[13] I have in mind here concerns raised by Karen Bennett that we will consider in the final chapter.

Chapter Two

Section A: Emergence

I have established four necessary conditions for closing the explanatory gap. First, our proposed theory must be physicalist; second, it must take qualia seriously; third, it must be reductive; fourth, it mustn't make use of unacceptable brute facts. We know that the first two of these are clearly required—physicalism, by hypothesis, and qualia. Yet the latter two conditions are suspect. It isn't obvious why a theory of mind would require phenomenal reduction to close the explanatory gap. Additionally, that a theory not posit a "bad" kind of brute fact is a vague requirement that threatens irreconcilable disagreement over what constitutes unacceptable bruteness. The fact is, the felt need for these requirements as expressed, for instance, by Levine, stems from a clearer problem that plagues the physicalist theories we've considered thus far.

What, exactly, is the problem with bruteness? Bruteness is required for any given theory: all theories will need to avail themselves of at least some brute facts, and bruteness is not by itself problematic. What of reduction? It's far from obvious that a proper physicalist theory of mind must be reductive. Nonreductive physicalism is a popular commitment. Individually, these two requirements seem iffy. However, the issues with the physicalist theories we have considered is not merely that they utilize brute facts or that they posit irreducible phenomenal qualities. Rather, the problem is that these theories posit brute identities between complex arrangements of matter and our high-level phenomenal states. Whatever else may be said for the views, at bare minimum this makes us uncomfortable. Levine has the intuition that the

problem is a problem about reduction, if only of an explanatory variety. But why? What is it about reduction that can make these theories satisfying, and what is it about its absence that's so jarring? The answer is that making phenomenal states brutely identical with complex physical states entails that those phenomenal states are emergent. I will provide a more precise definition shortly, but, for now, let us define emergence as follows:

> Emergence: *x* emerges from *y* if, and only if, *x* exists in virtue of *y* and exhibits properties not present in *y*.

On one understanding of emergence, emergent phenomena are, by necessity, irreducible. If mind turns out to be emergent in this way, I argue that the explanatory gap is unavoidable.

My suggestion here is that our hope for reductive explanation and the general discomfort with the type of bruteness articulated above is a gesture at the problem of emergence. Thus, I contend that if we are to solve the problem posed by the explanatory gap, we must have a theory that is nonemergentist. If this is right, then our four necessary conditions become three: we must have a theory that 1) is physicalist, 2) takes qualia seriously, and 3) is nonemergentist. However, there are a few things that need elucidation. First, we need to have a clear understanding of what it means for a phenomenon to be emergent. I must then show that emergence entails an explanatory gap. I also need to say something about the relationship between emergence and physicalism and the relationship between emergence and panpsychism, panpsychism being where I claim the solution lies.

Let's begin by talking about emergence and the explanatory gap. Emergence, under a certain understanding, is widespread and unproblematic. For example, liquidity appears to be

an emergent property of water. All that is meant here is that water exhibits some property that does not appear to exist in its constituents. While it may certainly be the case that water is liquid, it is not the case that its constituent molecules are themselves liquid. Water at the microphysical level does not exhibit liquidity: a single molecule of H_2O is neither solid, liquid, nor gas. The concept does not apply, for liquidity is to be found at the macroscopic level—it is of the relations that hold between water molecules. That properties can be emergent in this way is not controversial. More importantly, however, is the fact that the emergence of liquidity at the level of water is not even remotely mysterious. To quote Galen Strawson on the matter, "[t]he emergent character of liquidity relative to its non-liquid constituents does indeed seem shiningly easy to grasp. We can easily make intuitive sense of the idea that certain sorts of molecules are so constituted that they don't bind together in a tight lattice but slide past or off each other (in accordance with van de Waals molecular interaction laws) in a way that gives rise to — is — the phenomenon of liquidity" (2006: 13). A further fact about this type of emergence is that it allows for reduction. Liquidity can be explained completely in terms of the physical constituents of water along with their intrinsic and relational properties: "we can say that the phenomena of liquidity reduce without remainder to shape-size-mass-charge-etc. phenomena..." (Strawson, 2006: 13).

Emergence of this variety isn't troubling: the upper-level phenomena *transparently* depend upon the lower-level constituents. We can define this type of emergence as follows:

> Transparent Emergence: *x* transparently emerges from *y* if, and only if, *x* exists in virtue of *y* and exhibits properties not present in *y* that are reducible to the properties and relations of *y*.

Note also that transparent emergence makes no use of brute facts. The relationship between the phenomena at both levels is explicit: "[y]ou can get liquidity from non-liquid molecules as easily as you can get a cricket team from elevent things that are not cricket teams" (Strawson, 2006: 15). The emergence we're interested in is quite unlike transparent emergence. We can call the more troublesome kind of emergence 'brute emergence'. It is precisely this type of emergence that the theories we considered in Chapter 1 all made use of, and it is this type of emergence that necessarily results in an explanatory gap.

The brute emergence of experience is usually characterized as some type of dependence without reduction. We can define it as follows:

> Brute Emergence: *x* brutely emerges from *y* if, and only if, *x* exists in virtue of *y* and exhibits properties not present in *y* that are neither reducible to nor explainable by the properties and relations of *y*.

Theories that make use of brute emergence have a sort of mystical *feel* to them. There are two illustrations of just how baffling this kind of emergence can be that are worth briefly describing. The first makes use of the intuition that qualia are clearly present in the animal kingdom, though not everywhere. One might, then, believe that creatures such as humans, dolphins, chimps, and pigs have experiences, but creatures such as oysters, sponges, and jellyfish do not. But arranging the animals into a hierarchy of experientiality would be a Herculean task, and the number of creatures who would reside in the gray area would be enormous. More importantly, the differences between the creatures incapable of experience and those at the very bottom of the gray area would be minute. As though from nowhere, a creature that it's something to be like appears somewhere on the spectrum, and there's nothing

approximating an explanation to make sense of why qualia appear where they do. It also starts to look a bit funny that we would have the mental machinery to know where on the spectrum we can expect to find experience.

The second illustration, which can be found in David Skrbina's *Panpsychism in the West*, I find to be more powerful. The evolutionary chain ranges from where we stand as highly complex creatures all the way back to creatures of purest simplicity. The emergentist about experience is committed to the view that, at some point in our universe, qualia did not exist at all. Furthermore, when evolution first took hold, it created mindless creatures that did not feel the world at all. Evolution worked at a snail's pace, making tiny changes from parent to child, over millions of years, and it spent most of that time working on creatures that possessed nothing in the way of experience. Yet, the emergentist must hold that at some particular point in the timeline, an utterly unminded parent gave rise to a genuinely minded child for whom the world felt like something. From nothing came qualia, and the change was perhaps nothing more than a single gene. I believe Skrbina puts the point powerfully:

> ...at some crucial point in organic evolution, the first enminded creature appeared. That is, suddenly appeared. Some first select species—and indeed, some first individual organism— suddenly "felt" the world. Suddenly, the light bulb went on. [...] The miraculous nature of such an event is hard to overestimate. Mind came from that which was utterly devoid of mind. Enminded children came from utterly unminded parents. Mentality, subjectivity, qualia, suddenly appeared, like a bolt from the blue, having never existed in the known universe. This is brute emergence. (2007: 18)

One can, of course, claim that the minds of the very first creatures to possess them were simple, rudimentary, and quite unlike what we experience. But regardless of how impoverished the

mental experience of that being may have been, it remains remarkable that from whence there was nothing-it's-like came something-it's-like—from no experience came some experience.

This is all well and good, but it's time to elaborate what it means to say that a phenomenon is brutely emergent. Consider what Nagel says on the matter. According to him, "[a]ll properties of a complex system that are not relations between it and something else derive from the properties of its constituents and their effects on each other when so combined" (Nagel, 1979: 182). If we encounter properties of a complex system that cannot be explained by appeal to the constituents and their properties, we may call this an instance of emergence. However, Nagel goes on to say that this harmless kind of emergence "is an epistemological condition: it means that an observed feature of the system cannot be derived from the properties currently attributed to its constituents," which suggests that "either the system has further constituents of which we are not yet aware, or the constituents of which we are aware have further properties that we have not yet discovered." In other words, if we find a property of a complex system that we cannot explain, it must be the case that there are further facts about the *constituents* of that system that we are not privy to. It cannot, however, be the case that no such properties exist. That would make the property of the complex entity brutely emergent, or "truly emergent" in Nagel's terms, and "[t]here are no truly emergent properties of complex systems" (1979: 182). What we should take away from this is that any property of a complex system that is claimed to be irreducible to the properties of that system's constituents is brutely emergent. With this in mind, let's see why the theories we considered in Chapter 1 all utilize brute emergence.

Take behaviorism again. Behaviorism is, as we've said before, the claim that minds are behaviors—all mental states are sets of behavioral dispositions. However, everything in the universe exhibits some behavior or other. So, behaviorism must have it that there are some systems that behave and have minds, and there are some systems that behave and do not have minds. But the behavioral systems that possess minds do not have any new physical properties that are not present in behavioral systems that lack minds. Thus, the emergence of mind in the systems that possess it must be brute. For some unexplainable reason, some behavioral systems just give rise to mindedness. Reconsider the identity theory. It claims, roughly, that all phenomenal states are identical with some or other physical state. However, all objects in the universe have physical states. This means, as may sound familiar, that there are some physical objects that have mentality and others that lack it. The identity theorist does not grant mind to the idle rock. But the physical objects that possess mind are not fundamentally physically different from the ones that lack mind. To make the point a tad clearer: there's no new type of proton or electron present in enminded systems that is absent in unminded systems. So, once again, mind is brutely emergent; the mentality of enminded physical objects is not due to some special properties of the physical object itself. Rather, the identity is brute: mind brutely emerges from matter sometimes. Finally, let's turn our attention to functionalism once more. Functionalism claims that all phenomenal states are identical with some or other functional state. However, everything in the universe fulfills some kind of function—certainly the kind of function appealed to by functionalists. All objects realize some machine table or other, and all objects have internal states that mediate between inputs and outputs. So, as before,

functionalism must claim that only some of the functional states in the universe bring minds along, whereas others will lack minds altogether. But there is no new physical entity present in enminded systems that is lacking in unminded systems. Once again, this physicalist theory of mind appeals to brute emergence.

Brute emergence necessitates the existence of an explanatory gap. Moving forward, I shall use the terms 'brute emergence' and 'emergence' interchangeably, unless otherwise stated. For the sake of clarity, if *x* brutely emerges from *y*, then there is, ontologically, nothing that makes it the case that *x* emerges from *y*: *x* just does emerge from *y*, end of story. This is deeply problematic, and it is unexplainable in the strongest sense. In fact, Strawson takes this kind of emergence to be impossible, exclaiming that "*Emergence can't be brute*. It is built into the heart of the notion of emergence that emergence cannot be brute in the sense of there being absolutely no reason in the nature of things why the emerging thing is as it is (so that it is unintelligible even to God)" (2006: 18). Whether brute emergence is metaphysically possible is not something I'll address, but it is clear that brute emergence is, at minimum, epistemically troubling. If mind *emerges* from the physical, then there's nothing *about* the physical that makes it the case that mind should emerge from it. If there's nothing about the physical that makes it the case that mind emerges from it, then an explanation is straightforwardly impossible. A lack of the requisite explanation just is the explanatory gap. Therefore, if mind emerges from the physical, then there must be an explanatory gap.

There are those who claim that brute emergence is precisely what is necessary to explain the existence of mind in the physical world, but this seems like an act of desperation in

the face of no other good alternative. Appealing to brute emergence to explain mind requires too much. Strawon puts the requirement in vivid terms: "One problem is that brute emergence is by definition a miracle every time it occurs, for it is true by hypothesis that in brute emergence there is absolutely nothing about X, the emerged-from, in virtue of which Y, the emerger, emerges from it" (2006: 18). However, that emergence requires a true miracle is not a claim I need defend. I need only accept that the physicalist theories we have considered (and those that avail themselves of similar tactics) are all emergentist theories. For our purposes, the weaker claim that emergence entails an explanatory gap suffices. As such, if we're to have any hope of closing the gap, we're going to need a nonemergentist theory.

Now, I don't believe that emergence is incompatible with physicalism. Two of the three necessary conditions that I have laid out could fail to be satisfied by a truly physicalist theory—we need these three requirements to be met so that we may have a *satisfying* physicalist theory that does not suffer from an explanatory gap. Some of what Nagel says suggests that physicalism and emergentism are incompatible, but this is because he seems committed to the view that there are psychophysical laws and that causation, in proper physicalism, is a form of necessitation.[14] Briefly, if mental states brutely emerge from physical states, then they possess properties that the physical states they depend upon do not possess. If causation is a form of necessitation, then all causes necessitate some particular effect. Now assume that a mental state brutely emerges from a physical state. The mental state and the physical state upon which it depends will differ in their properties and thereby have different causal profiles, but they must

[14] He explicitly rejects causation as mere correlation.

play in to all of the same causal interactions. Thus, when I stub my toe, we want to say that this necessitates my pain, but there's nothing about the world that makes this necessitation the case. If necessary causal connections are required for physicalism, then any case of emergence will violate physical laws and thereby violate physicalism. Or so I believe the reasoning goes. But physicalism is perfectly compatible with a Lewisian view of causation, and this type of causation is more amenable to the possibility of emergence. Strawson seems to share Nagel's sentiment, claiming that physical stuff, given the truth of what he calls 'real physicalism', when "put together in the way in which it is put together in brains like ours [literally is] experience like ours" (2006: 9). His point is that, in order for a theory to be truly physicalist, it must allow for some kind of reductive explanation, effectively barring emergence. But this isn't quite right. While I agree with Nagel and Strawson on the major point about the unworkability of brute emergence, one can certainly be a physicalist and an emergentist. All one really needs to be a physicalist is to subscribe to the view that the world is made up of only physical stuff. To consider an example, it would be possible to live in a purely physical world where placing six electrons within one micrometer of one another in a hexagonal arrangement results in the group exhibiting a powerful positive charge. The positive charge brutely emerges from the placement of the electrons, none of the components exhibit positive charge, and nothing nonphysical has occurred.

If emergence and physicalism were incompatible, then behaviorism, the identity theory, and functionalism would be nonphysicalist theories. While I feel the intuitive pull of this claim, I also believe that it is inexpensive to claim that the phenomenal is in some sense

physical. It's not clear what this would amount to (and doing so by itself solves no problems), but given that we do not have a clear definition of 'physical', there's no reason to believe that the phenomenal cannot be construed as physical. Indeed, this is precisely the kind of move that any physicalist theory of mind will have to make. I, myself, will be providing a clear place for the phenomenal within the physical world as a respectably physical entity in its own right. This is, I take it, all one needs to be a physicalist, though the plausibility of claiming the mental as physical comes in varying degrees depending on the theory. All the same, once one allows for emergence, one is condemned to dealing with the explanatory gap, and closing it is a hopeless task: no amount of scientific poking and prodding will solve the problem.

So, we've established that while physicalism is indeed compatible with emergence, physicalist theories that rely on it will by necessity be left with an unbridgeable gap in explanation. The solution, I submit, is panpsychism. Sort of. Panpsychist theories are not automatically nonemergent. As we will see shortly, it is possible to be a panpsychist and also be left with an unbridgeable explanatory gap. Furthermore, it isn't the case that panpsychist theories will automatically be physicalist theories. Nonetheless, panpsychism presents our only way out. While panpsychism may not automatically grant us a nonemergentist theory, *only* panpsychism can. Of all of the types of theories that aim to be physicalist and take qualia seriously, panpsychism is the only one that can deliver on nonemergence. I will provide an argument for this claim in what follows.

I must now take on the task of characterizing panpsychism. I'll provide a broad understanding of what panpsychism is, and I will offer a rough taxonomy that'll suit our

purposes. Afterward, we'll eliminate the versions of the theory that are not up to the task—those that cannot meet all three of our necessary conditions. By the end of this chapter, we'll have a theory that shows real promise, though we'll then need to take on the task in Chapter 3 of justifying it as a truly physicalist theory.

Section B: Panpsychism

Panpsychism, while being one of the oldest theories of mind, is poorly understood. It is occasionally construed as the view that everything is conscious. The term 'consciousness', however, "is highly anthropocentric, and its meaning is closely associated with specifically human states" (Skrbina, 2007: 8). So, one might derisively claim that the panpsychist believes that a pair of socks is conscious.[15] It is unlikely that anyone holds this view. Panpsychism might also be mistakenly believed to be the claim that all entities possess thought, which is, again, considered a very human-centric mental phenomenon. Now, while both of these options would count as panpsychist, they're deeply uncharitable characterizations. Most panpsychists wouldn't believe anything approximating these claims. Furthermore, panpsychism is less of a theory and more of an umbrella for a diverse set of theories with a common theme. As such, panpsychism, broadly understood, has less concrete commitments.

So let's clarify how panpsychism should be understood. Skrbina offers a useful characterization:

> I should note here that panpsychism doesn't entail that every conceivable entity possesses mind. For example, valid panpsychist theories may exclude composite or

[15] Karen Bennett (2005) comes close to saying something like this.

> collective entities, such as piles of sand, or tables and chairs. They may exclude physical ultimates such as atoms—or they may include *only* physical ultimates. They may include matter but exclude various forms of energy. They may exclude conceptual or logical entities, such as numbers. I will therefore interpret panpsychism in the soft sense: that mind is very widespread, is nearly universal in extent, and crosses deeply into the inorganic realm. The precise extent of mind depends on the particular theory at hand. (Skrbina, 2007: 3)

Panpsychism, as characterized above, is a commitment to the belief that mentality, in some degree, is almost universally present. However, this seems to fall short of what one might think *pan*psychism should be.[16] The important claim that panpsychism makes for our purposes, however, and the claim that many find to be surprising, is that mentality is far more widespread than we commonly believe. Matter that is often labelled 'inanimate' possesses some degree of mind. Notice that terms such as 'inanimate' are exceedingly difficult to define, and any definition offered seems subject to counterexamples. Given the sheer breadth of what gets to count as panpsychism, there are many theories that we can construct that can rightly claim the name. Most of those theories will fail to meet the requirements for closing the explanatory gap. To work through them all would require too much space and be unnecessary. I will instead offer a rough taxonomy that will allow us to find our way to the viable theories quickly.

We can split the panpsychist field into two broad categories: macro and micro views. Let's begin with the macro views. Macro panpsychist theories are those panpsychist theories which posit mentality at all (or most) concrete levels of existence. These views will claim that

[16] I won't concern myself in this dissertation with the terminological issue. Most philosophers who concern themselves with panpsychism seem to accept calling theories that fall just short of ascribing mentality to all things 'panpsychist'. Nothing important hangs on this, and I'm happy to follow standard practice here. I will say that the theory that I put forth is one I believe to be truly *pan*psychist.

things like atoms, microorganisms, the Earth and Sun, and the cosmos as a whole all possess minds simultaneously. It isn't necessary that they all grant mind in this manner, but there is a common theme: large-scale entities possess mind, and that mind is not composed of smaller minds. In fact, the physical components of these large-scale entities can possess minds at the same time. The minds of these smaller entities are not part of the greater mind; they do not merge and they needn't interact. One such view articulated by Skrbina is that of Gustav Flechner, where Flechner claims that the world is "composed of a hierarchy of minds or souls [with] souls 'below' us in the plants [and] 'above' us in the Earth, the stars, and the universe as a whole" (2007: 146). Indeed, that there exists a kind of universal world-soul has been a surprisingly commonly held view throughout philosophical history (Skrbina, 2007). We, of course, are part of the cosmos, and we would have souls on this view as well. But our souls are not part of the soul of the cosmos, even if our matter is part of its matter.

Micro panpsychist views are those that posit mentality at the fundamental level of reality. These types of views posit phenomenal (or protophenomenal) properties or entities somewhere around the fundamental level. There are, broadly, two types of micro views. There are those that posit phenomenal properties just above the absolute fundamental level, and there are those that claim that the most fundamental entities themselves possess a degree of mentality. The former views might, for instance, grant molecules experience but deny it to atoms, or perhaps subatomic particles; the latter ones will place it wherever reality bottoms out. On the micro views, there are indeed still conscious entities at the higher levels, but they have the lower-level mental stuff as components—these minds are somehow built up from simpler

mentality, bottoming out around the fundamental level. These views are tasked with providing an explanation for how this is supposed to work, and this explanation should presumably explain which types of higher-level entities possess minds. One such view that we'll see again later is that the fundamental level of reality is composed of 'mind-stuff', a psychical, simple kind of stuff that can explain how minds like ours come to be.[17] Micro panpsychist views can vary widely—some going as far as positing the existence of psychic atoms that serve as psychical counterparts to physical atoms (Skrbina, 2007: 207). We won't consider this type of view, however, as it would clearly fail to be physicalist.

Panpsychist theories, as previously stated, can be emergentist and thereby fail to meet one of our criteria. All macro views of panpsychism make use of emergence. For instance, if all objects have souls, then these souls will undeniably have to brutely emerge from whatever gets to count as an object, and the base components of these objects will either be devoid of soul-stuff or, if they possess soul-stuff, that soul-stuff will not play into an explanation of how the greater soul gets to be possessed by the composite object. Such a view would, of course, also fail to count as physicalist, though we needn't rely on souls to get the point about emergence across. All macro views will face this difficulty. Even without souls, the macro panpsychist is forced to posit mentality at macroscopic levels without any mechanism for reduction. Not only do these minds not reduce to their physical constituents, they cannot even reduce to the mental properties of their physical constituents. If they could reduce in this manner, then we would be dealing with a micro panpsychist view.

---

[17] This view is defended by William Clifford (1878), who coined the term 'mind-stuff'.

Micro views are the only ones that can avoid emergence, then. Note, by placing the phenomenal at the base level of reality, micro views make all of the mentality at that level brute. To be specific, any phenomenal property that a fundamental entity possesses is related to that fundamental entity brutely. The phenomenal character of an electron is not subject to explanation. Bruteness at this level is typically unproblematic (Strawson, 2006; Levine, 1983). We expect some things to be brute, and that a phenomenal quality be brute at the fundamental level, whatever that may mean, should in itself be no more problematic than negative charge being a brute property of electrons. Placing the phenomenal at the fundamental level does not guarantee a nonemergentist theory. For instance, Philip Goff articulates a form of micro panpsychism that takes emergence as necessary in his *Galileo's Error* (2019: 164-72). However, micro views provide the promise of avoiding emergence. If the phenomenology we are familiar with can somehow reduce to more basic phenomenology of fundamental entities, then we will have successfully avoided the problems of emergence.

Between the macro and micro views, only the micro views have a shot at getting the job done. At this point, this is nothing more than promissory. The promise: I will deliver a micro panpsychist theory that can properly count as physicalist and nonemergentist. What we have now is a very imprecise understanding of a kind of panpsychist theory that is compatible with nonemergence and may get to count as physicalist upon being further fleshed out. The only thing we really know is that the theory must posit mentality of some variety at the lowest level of reality. The remainder of this chapter is dedicated to constructing to a high degree of precision what our micro panpsychist theory must look like in order to be successful.

Section C: Two Types of Micro Theories

Micro panpsychist views, as we've said, place mentality at or close to the fundamental level of reality. Now is a good time to rule out the views that fall short of placing experience at the very lowest level of existence. If the fundamental components of matter are utterly without phenomenal character, then phenomenology becomes present once those fundamental components either combine or interact in some way. For instance, this might happen when unminded quarks come together to form minded protons. This won't do, for it will invariably give rise the explanatory gap once again for familiar reasons. As with the identity theory, we would be relating some kind of mentality to a complex entity. Though the degree of complexity involved is significantly lesser, the problem remains. That unminded materials come together to form minded materials is kept brute, and so the same issues as before arise once again. It is for this reason that we cannot make use of micro views that fail to place mentality at the very bottom. The only micro views we have available, then, are those that do place the mental at the fundamental.

I have thus far remained silent on what it means for mentality to exist at the lowest level of reality. The inner lives of, say, electrons (which will be our go-to example) remain unspecified. What it is like to be an electron is bound to be beyond our imaginative grasp, but I believe it lies within our cognitive reach to offer some kind of description: the question of the nature of fundamental phenomenology is answerable in the abstract—whether it is experience of a complex kind or simple kind, vivid or dim, similar or dissimilar to our own, etc. I have said nothing about this yet. For the sake of differentiating between the experiences of minds like

ours and the mentality of fundamental entities, I will use terms such as 'protophenomenal' and 'protophenomena', and I will occasionally make use of the prefix 'proto-' when talking about the mental character of things like electrons. For the moment, these terms will serve only the purpose of differentiating between the phenomenal experiences we are familiar with and those of what are often called 'simples' or 'ultimates'. Whether the experiences of such simples bear any of the marks of the mental is a question we'll take up later.

We know already that fundamental particles exhibit physical properties, and I am now claiming that these physical properties do not provide an exhaustive description of the microphysical world. Fundamental entities like electrons have physical properties, such as negative charge and spin. Our claim—the claim of the micro panpsychist views still available to us—is that we can find protophenoma at this level as well. So, electrons, in addition to exhibiting charge and spin, are also the bearers of protophenomenal qualities. Such views have been defended by philosophers such as William Clifford, who claimed that a basic molecule of inorganic matter, while not possessing full-blown mind, "possesses a small piece of mind-stuff" (1878: 65). These fundamental entities must somehow have both physical and protophenomenal properties. I must now answer the question: how are these types of properties related? There are two ways that I can make sense of this relationship.

The first option is to claim that electrons bear two distinct kinds of properties: physical and protophenomenal. This is akin to the commitments of the property dualist, a major difference being that most property dualists do not take themselves to be panpsychists. As a side note, being a property dualist about the properties of fundamental entities strongly

suggests that one should be a panpsychist. There is no clear reason to believe that the electrons present in brains have protophenomenal properties but not those that are present in chairs and tables. Without an elaborate story, the threat of inconsistency looms over such a view.[18] Similarities aside, the option we are entertaining here does not amount to property dualism. Many property dualists consider themselves straightforwardly not physicalists. The phenomenal or protophenomenal properties they may make use of are not in any sense physical—they are additional nonphysical properties. The option we're considering is meant to be construed in only physicalist terms. The views I have in mind are those of the nonreductive physicalists. I take nonreductive physicalists to posit exactly this: there are phenomenal and physical properties which are different from one another, and the phenomenal properties supervene on the physical or are otherwise highly dependent. The dependence relation is intended to be strong enough for such views to be properly physicalist. As I see it, there is nothing to stop this kind of supervenience relation from holding all the way down to the fundamental level such that the protophenomenal supervenes on physical properties. Furthermore, the protophenomenal (or phenomenal) and physical are importantly different, as any nonreductive physicalist will stress. Such a view could be called 'nonreductive panpsychism': the view that the fundamental entities of the world have distinct protophenomenal and physical properties, and the protophenomenal properties supervene on the physical.

[18] Most property dualists are likely to posit the existence of phenomenal properties only at a macroscopic level, though this does seem to raise the issues mentioned in Chapter 1.

The second option we have for relating protophenomena and physical properties is to identify them. In looser words, the mental and physical somehow turn out to be the same thing: protophenomenal properties just are physical properties. There are a number of things this might mean. For instance, Nagel suggests reducing the mental and phenomenal to a common base, which would "have the advantage of explaining how there could be necessary causal connexions in either direction" and would offer other advantages (1979: 184). This is, of course, only one possibility.

On the face of it, the first option is the more intuitive. That the phenomenal is supervenient upon and does not reduce to the physical is certainly prevalent in the literature and enjoys a wide breadth of support.[19] I take it that its panpsychist cousin will be similarly seen as the more viable of the two views we have on offer. Nonreductive panpsychism certainly takes qualia seriously, and it apparently avoids issues of emergence. However, I will argue that this view is not truly physicalist. The argument against the nonreductive panpsychist equally affects standard nonreductive physicalism. Indeed, I will be appealing to Jaegwon Kim's familiar causal exclusion argument in objecting to the view. By rejecting the first option, we are left with the second. But much remains to be said about what it means for the protophenomenal to be identical with the physical. Without an understanding of what exactly is meant by 'protophenomenal', it isn't clear that the view can avoid the explanatory gap. Indeed, this is the objection raised by Bennett in her paper "Why I Am Not a Dualist," and it is where I will focus my efforts toward the end of the chapter. Additionally, it is far from clear

[19] E.g., see Bennett (2003), Davidson (1970), Fodor (1974), Melnyk (2003), Putnam (1990), and Yablo (1987). At least some of Chalmers' (2003) Type-B materialists will have to keep the two distinct.

what it would mean for the protophenomenal and the physical to be identical. Making sense of this is of vital importance. If I cannot make the claim sensible, then the theory won't get off the ground. So, we'll need a characterization that is plausible and that is prepared to offer a meaningful response to the problem posed by the explanatory gap. To this I dedicate Chapter 3. For now, let us draw our attention to the first option: nonreductive panpsychism.

Section D: Nonreductive Panpsychism

According to the view currently at issue, there are two distinct types of property: the physical and the protophenomenal. Sticking to the convention established in the previous section, we can call it 'nonreductive panpsychism'. Before proceeding, it's important to say something about the relation between nonreductive panpsychism and nonreductive physicalism. In what follows, I will be arguing that nonreductive panpsychism is not a truly physicalist theory. In doing so, I will be appealing to standard arguments against nonreductive physicalism. Nonreductive panpsychism and nonreductive physicalism share much in common, but they are markedly different theories that make sharply different claims about the world. There are at least two main differences. First, the nonreductive physicalist is likely to posit phenomenal properties at the macroscopic level as properties of objects like us, whereas the nonreductive panpsychist will posit these properties at the microscopic level. Second, and most obviously, the nonreductive panpsychist grants mentality to all fundamental entities, whereas the nonreductive physicalist is more 'chauvinistic', to borrow the term from Block, about what gets to have a mind. Nonetheless, the nonreductive physicalist and the nonreductive

panpsychist each claim the same kind of relationship holds between the mental and the physical. It is precisely this relationship that, I argue, prevents the view from being truly physicalist. Hence, for our purposes, these two theories stand and fall together. The objections that I consider and the defenses that each theory can offer to those objections are all the same. Given this, every claim that I make in what follows, unless otherwise specified, targets both theories. I will jump between talking about the nonreductive physicalist and the nonreductive panpsychist as appropriate.

Now, given that physical properties on the view on offer are importantly distinct from protophenomenal properties, what reason do we have to claim that the theory is physicalist? There must be something about the relationship between the two that can justify the label. The view is not, by itself, obviously physicalist. In order for the nonreductive physicalist/panpsychist to be a true physicalist, there must be some sort of special relationship that holds between the mental and physical that permits the mental to play a meaningful role in the physical world. In other words, the mental must earn its keep. The standard relation appealed to in order to achieve this end is supervenience. Supervenience is typically seen as a dependence relation whereby that which supervenes is dependent upon its supervenience base. Intuitively, it's the mental that must supervene on the physical. Given that there are different types of supervenience relations that vary in strength, it isn't immediately obvious how tightly the supervenience relation between mental and physical properties should be construed. The tightest available supervenience relationship is logical supervenience, which identifies the

supervenience base with the supervening class. More will be said momentarily, but let us begin with a definition which I take in paraphrased form from Chalmers (1996: 35):

> *Logical Supervenience*: A-properties logically supervene on B-properties if, and only if, it is logically impossible for two things to possess identical B-properties and yet differ in their A-properties.

As a straightforward example, the shape of an object logically supervenes on the arrangement of its parts such that no two objects in any world (or across worlds) can have identical structures with differing shapes. Chalmers notes that the fact that the relation between the A-properties and B-properties is one of logical supervenience does not require that we be able to deduce one from the other. It suffices that the supervenience base necessitates its supervening properties in the sense that all there is to the existence of the supervening properties is that the supervenience base is present (Chalmers, 1996: 36). To borrow the example Chalmers uses, the biological facts of a world logically supervene on its physical facts.[20] Once all of the physical facts of the world are settled, all of the biological facts are as well. Two worlds cannot be physically identical and yet biologically distinct. Were protophenomena to logically supervene upon physical properties, then, it would be impossible to have two worlds (or beings) identical in their B-properties who differ in their A-properties.

This is clearly a supervenience relation nonreductionists cannot help themselves to. If mind logically supervenes upon body, then mind and body are identical. Were they not identical, it would be logically possible to have qualitatively identical bodies that differ

[20] Chalmers adds in some further restrictions to the supervenience relation so as to disallow the possibility of immaterial souls in some worlds barring the relation from holding. I ignore these details here.

mentally, which is just a rejection of the logical supervenience relation. The nonreductive physicalist doesn't care for this brand of supervenience anyway, as there is interest in keeping the mental irreducible and autonomous (Kim, 1989: 32). They will have to appeal to a weaker supervenience relation. One such weaker supervenience relation is aptly named 'weak supervenience'. I borrow the following definition in paraphrased form from Kim (1987):

> *Weak Supervenience*: A-properties weakly supervene on B-properties if and only if it is impossible for two things to possess identical B-properties and yet differ in their A-properties *in a given world*.

Weak supervenience is concerned solely with what occurs within a given world. One might think, for instance, that the fragility of a given glass weakly supervenes on its given microstructural properties. If this is right, then any two glasses with the same microstructural properties within a given world will need to be identically fragile. So it must be with minds if we claim weak supervenience. If protophenomenal properties weakly supervene on physical properties, then two physically identical entities must also possess identical protophenomenal properties within the same world. This importantly allows for the existence of distinct worlds where physical duplicates of objects in this world vary protophenomenally.

There is one final type of supervenience that I believe it will be beneficial to articulate for the present discussion, and that is global supervenience. Global supervenience, unlike weak supervenience, is specifically about entire worlds.[21] Following Kim, we can take the following paraphrased definition of the term (1989: 41-2):

---

[21] Notice that logical supervenience can be about worlds or individuals. The two are not opposed, though they are distinct concepts.

> *Global Supervenience*: Worlds that possess identical B-properties cannot differ in their A-properties.

Global supervenience, then, requires that worlds that are duplicates in regards to their B-properties must be duplicates in regards to their A-properties. To use the example of biology again, two completely physically identical worlds will be biologically identical. Furthermore, if protophenomenal properties globally supervene on physical properties, then two physically identical worlds will be protophenomenally identical. Notice that this type of supervenience is stronger than weak supervenience but weaker than logical supervenience. Unlike weak supervenience, global supervenience can make cross-world claims, but unlike logical supervenience, global supervenience cannot make claims about individuals across worlds, only the worlds themselves. I've opted for these three types of supervenience precisely because they offer a nice spectrum of options, though the list is by no means exhaustive. What matters most for our purposes is that we have weaker alternatives to logical supervenience, as the logical variety will be unavailable to the nonreductive panpsychist.

With these definitions in mind, we can now move on to what the nonreductive panpsychist must claim and where the theory goes wrong. The nonreductive panpsychist will have to opt for something like weak supervenience. Furthermore, the theory must justify itself as properly physicalist. Mere supervenience will not suffice; in order for mental properties to be appropriately physicalist, they must have some causal work to do—they must play some causal role in the world. In Kim's words, we "had better find some real causal work for [our] mental properties" if keeping them around is to be more than "a token gesture" (1989: 43). A physical

property that does absolutely nothing at all (and can do nothing at all) does not look like a physical property. Indeed, it would be physical in name only. Mental properties must be causally efficacious, but this is precisely what Kim claims cannot happen: if the mental supervenes on the physical without reduction, then it cannot play a causal role in the world. To defend the point, Kim utilizes his well-known causal exclusion argument (1989: 44-5). Let's run through it.

First, we must acknowledge three principles the nonreductive physicalist must accept. The first is the principle of Causal Closure, which states "this: *any physical event that has a cause at time t has a [sufficient] physical cause at time t*" (Kim, 1989: 43). The intuition here is straightforward: if there are any nonphysical causes of physical events, then the world is not physicalist. This, by itself, does not bar there being other causes, but all physical events *must* have a sufficient physical cause. The second principle disallows rampant overdetermination. We can call this the 'No Overdetermination' principle as it's often called, and it states that while there can certainly exist instances of genuine overdetermination, they are rare. Thus, our theories had better not have the consequence that there is rampant, genuine overdetermination. The third and final principle is that of Supervenience: the mental (be it phenomenal or protophenomenal) supervenes on the physical. In what follows, we will be specifically concerned with weak supervenience, though any kind weaker than logical that gets the job done will do. Kim then asks that we imagine some mental event M_1 causing some physical event P_2. By the principle of Supervenience, M_1 must have supervenience base P_1. By Causal Closure, P_2 must have a sufficient physical cause, in this case P_1. So, it seems that we

have two candidate causes of P_2: P_1 and M_1. They cannot be partial causes that are together sufficient for P_2. If they were, then that would mean that had P_1 happened by itself, P_2 would not have occurred, which would violate Causal Closure.[22] If both P_1 and M_1 are individually sufficient causes, then this is a violation of the No Overdetermination principle. Furthermore, this would also be a violation of Causal Closure, for it would be to say that M_1 without P_1 would have caused P_2. Therefore, it must be the case that P_1 is the true cause of P_2, so M_1 is causally excluded.[23]

It seems that for every mental event, there will be some supervenience base that is physically sufficient to bring about the relevant physical cause. So the mental is left without a causal role to play. Bennett emphasizes the point: even if protophenomena were capable of taking an active role in the causal world, "even if they *are* perfectly suited to causing things, there is nothing around for them to cause" (2003: 471). If the mental is never causally efficacious—if it cannot play a *causal* role in the world—then it plays *no* role in the physical world. Nonreductive physicalism, and thereby nonreductive panpsychism, is not truly physicalist. The view, however, is not without defense.

Kim's argument hangs on the three principles previously articulated: Causal Closure, No Overdetermination, and Supervenience. To resist his argument, at least one principle must be either denied or shown not to apply in the case of mental causation. We cannot reject Causal

---

[22] We are assuming that P_1 is the complete physical event that caused P_2, not that it had some further physical event helping it (on which perhaps M_1 was supervening).

[23] One might want to claim that the mental is still causally present in the world, but only as an effect and never as a cause. Thus, perhaps M_1 is the effect of some previous P_0. However, every mental event will supervene on some physical event, and there will be no reason to claim that the prior physical event caused the mental event directly rather than through causing the mental event's supervenience base.

Closure or Supervenience. Causal Closure will be necessary for any physicalist theory, and Supervenience is a central tenet of the nonreductive project. This leaves No Overdetermination as the only potential target. In order to save the causal efficacy of the mental, it must be the case either that the No Overdetermination principle is unwarranted or that it simply does not apply. Let's now briefly consider two arguments that have been offered in defense of nonreductive physicalism that target the No Overdetermination principle. These arguments are meant to show that the causal efficacy of the mental is exempt from counting as overdetermining its effects. In other words, they do not claim that the principle doesn't hold, but rather that mental events cannot be ruled inefficacious on grounds of violating No Overdetermination. I will show, *pace* these arguments, that both are successful only if they make use of reduction. Absent reduction, Kim's causal exclusion argument goes through, rendering nonreductive views nonphysicalist.

The first defense is offered by Bennett (2003). She begins with a few thoughts on overdetermination. We believe that we have a genuine instance of overdetermination, according to Bennett, only if two counterfactuals hold. Since overdetermination occurs when there are two sufficient causes for an event, it must be the case that had either cause occurred without the other, the event still would have transpired (these being the two counterfactuals). In an effort to save the nonreductive physicalist, Bennett believes that she can defend at least one of two claims when it comes to mental causation. The first claim is that the counterfactuals, while true, are only vacuously so, because at least one has a necessarily false antecedent. Furthermore, in order to have a genuine instance of overdetermination, it must be the case that the

counterfactuals are nonvacuously true. The second claim is that one of the counterfactuals is actually false, and so we are not dealing with an instance of overdetermination.

Before continuing, a quick note on the strength of the requisite supervenience relation Bennett will make use of. Bennett's arguments depend on a very tight supervenience relation between the mental and physical, though she doesn't specify which particular relation she has in mind. It's worth making clear here, by the way, that Bennett doesn't necessarily hold these views; she's offering these arguments as an act of charity to those who do. Anyway, the idea is that because the mental *supervenes* on the physical, we are not really dealing with independent causes, and so overdetermination shouldn't be a concern: the mental can be causally efficacious along with the physical without violating the No Overdetermination principle.

Let's begin with her argument in defense of her first claim. Consider the possibility that one of the overdetermination counterfactuals is vacuously true. First, it's quite unlikely that the mental state of pain supervenes on the firing of C-fibers. After all, claims Bennett, if we fire some C-fibers in a petri dish, it's not likely that there's anything that's feeling pain. So, pain is likelier to supervene on a more complex physical state, perhaps including the entire state of one's body plus some external facts. Either way, all we really need is that pain supervenes on some more complex physical state for her argument to get across. Let's suppose that pain weakly supervenes on the complex of all relevant physical properties, both internal and external, to the production of a physical pain response. So now, keeping in mind that pain weakly supervenes on this complex physical state, let's consider our overdetermination counterfactuals.

First: If the C-fiber complex were to occur without the pain, the yelp would still have occurred. Is this true? Yes, but only vacuously so, as it is impossible for the C-fiber complex to occur without its supervenient pain. What of the other counterfactual? Here it is: If the pain were to happen without the C-fiber complex, the yelp would still have occurred. Is this counterfactual true? Yes, but only because the antecedent is impossible, says Bennett. After all, since the pain supervenes on the C-fiber complex, the removal of one necessitates the removal of the other, granting us a conditional with a necessarily false antecedent. But this isn't quite right, Bennett admits. It's possible to remove the C-fiber complex and keep the mental state of pain: if some replacement occurs such that a new complex upon which pain supervenes is introduced, then the pain exists without the C-fiber complex.[24] Is this a problem? Maybe. If we believe that both counterfactuals must be non-vacuously true for genuine overdetermination, then we really need only the first counterfactual above to be vacuously true, and it certainly seems to be. If, however, we believe that overdetermination requires only one of the counterfactuals to be non-vacuously true, then there may be a problem with the latter counterfactual. I'm sympathetic, as is Bennett, to the possibility that both counterfactuals must be nonvacuously true in order for something to count as a genuine instance of overdetermination. If it really is only vacuously true that the C-fiber complex without the pain brings about the pain response for the sole reason that we cannot have the C-fiber complex

[24] I think we can bar this on the grounds that what we're interested in finding out is whether *this* pain was causally efficacious. Removal of the supervenience base of this pain will remove this pain by necessity, even if we could replace it with some other pain of the same type (by finding some other adequate supervenience base). Thus, this would still be a good test for the causal efficacy of some particular pain. If this one pain wasn't efficacious, the result would generalize to all instances of pain.

without the pain, then I am happy to grant that we cannot rule the pain causally inefficacious on grounds of overdetermination. In this case, the effect is not overdetermined. Granting that both overdetermination counterfactuals must be non-vacuously true in order to have a genuine instance of overdetermination makes my job harder, so we'll run with it. We'll grant for the time being that it is plausible that one of the overdetermination counterfactuals isn't the right kind of true.

Consider now the alternative possibility that one of the overdetermination counterfactuals is false. Bennett considers that we may not want to appeal to the entire C-fiber complex in writing up our counterfactuals. Perhaps that's building too much into them. Here's the counterfactual: were the C-fibers to fire without the pain, the yelp would have occurred. Is this true? Bennett says no. The worlds in which we have C-fibers firing without pain present are worlds in which C-fibers are perhaps firing within the confines of a petri dish. As such, we've removed the C-fibers from the pain, but doing so prevented the yelp. So, once again, we have reason to believe that the causal efficacy of the mental state does not result in overdetermination.

If the supervenience of the mental on the physical makes it impossible to divorce the two such that one of the counterfactuals turns out either false or vacuously true, then we cannot dismiss the causal efficacy of the mental on the basis of the No Overdetermination principle. However, what has not been shown is that this supervenience relation ties these two together so tightly. Just how tightly is Bennett taking this supervenience relation to be? Consider the fact that counterfactuals are often talked about in terms of what happens at other

worlds. So, to say that the C-fiber complex cannot occur without the pain is to say something about what other worlds are like. It has to be the case that we won't find instances of C-fiber complexes without pain. We've been considering the supervenience claim that the mental weakly supervenes on the physical. All this entails is that every instance of the C-fiber complex *in our world* brings pain along with it. This leaves it entirely open what happens in other worlds. It is consistent with weak mind-body supervenience that there be a physical duplicate of our world in which the C-fiber complex exists without pain coming into the picture. If both worlds are physical duplicates of one another, then the counterfactual is not only true, but non-vacuously so, at least if the relation is weak supervenience. But this needn't be the relation we use.

The supervenience relation Bennett is appealing to is remarkably strong, certainly stronger than weak supervenience. Indeed, in talking about the tightness of the relationship, she says that "the idea that it is metaphysically necessary that one of the causes occurs whenever the other does gives some content to the often-heard idea that despite not being identical, the mental and physical causes are not exactly *distinct*, either" (Bennett, 2003: 480). This is why the overdetermination counterfactuals won't turn out the way Kim wants them to, as she says in the sentence that immediately follows: "it also means that there is a sense in which one of the overdetermination counterfactuals is not quite up for discussion—you cannot quite ask what would happen if the one occurred without the other if it just *can't* occur without the other." To claim, then, that the antecedent is *impossible* is to claim that there are *no worlds* in which the antecedent holds. But as we've seen supervenience does not by definition relate two things in

such an inseparable way. In making a separate point, Bennett says that "[t]hough there may not be any souls *here*, there are worlds in which there are, and in those worlds things can have [mental property] M without any physical properties at all" (2003: 484). But of course, if such a world could exist, then the supervenience relation between mind and matter would have to be much weaker than she characterizes it, for that would be a world in which the mental could change without there being any changes in the physical. This kind of supervenience could perhaps be global supervenience, such that "any world that is just like this world in all physical details must be just like it in all psychological respects as well" (Kim, 1989: 41). But global supervenience, as Kim rightly says, is consistent with the existence of two almost physically identical worlds, but for the placement of a single electron, that are nothing alike mentally. If the relation between the mental and physical were a global supervenience relation, then for any one mental event, anything could serve as its physical base consistently with this relation, making the overdetermination counterfactual true once again. Of course, no nonreductive physicalist should use global supervenience in theorizing.

Bennett acknowledges the sheer strength of the supervenience claim needed. Indeed, she says that anyone who wishes to utilize her argument "has to deny the genuine possibility of zombie worlds. If there is a minimal physical duplicate of our world that is devoid of mentality (or, at least, is devoid of consciousness), then neither of the solutions I have suggested gets off the ground" (Bennett, 2003: 491). To claim that there are no zombie worlds, I believe our only viable option is reduction. We will have to claim that the protophenomenal logically supervenes on the physical. The alternative would be to dig our heels in, claim that

protophenomenal and physical properties are distinct even though they are present together in all worlds, and so dismiss the suggested identification without argument. I'll say more on the plausibility of this move in a bit. For now, let's consider a related but different defense.

Lawrence Shapiro also attempts to protect the causal efficacy of the mental from Kim's causal exclusion argument. Shapiro begins his defense by, like Bennett, stressing the fact that the mental supervenes on the physical and is thereby not independent and not subject to overdetermination concerns. "Overdetermination, as Kim recognizes, requires the existence of independent sufficient causes. Yet, the appearance of two independent sufficient causes of [some physical event] P* [...] is nothing *more* than an appearance. Given that [a mental event] M supervenes on [a physical event] P, M is not independent of P" (Shapiro, 2010: 595-6). Shapiro is right to say that the two are not independent, though I believe it is a mistake to claim that Kim needs independence for his argument to work. All he really needs is the distinctness claim that the nonreductive physicalist is committed to. The point is, regardless of the relation, there are *two* things, not one, that are involved in the causal event.

Shapiro construes Kim as making a claim about probability. In considering the possibility of P existing without M, Kim, according to Shapiro, is concluding that the effect P* is still guaranteed to occur. The claim is supposed to be that the probability of P without M causing P* is 1. On the basis of this claim, Kim causally excludes M, for the removal of M has done nothing to lower the probability of P* coming about. But, says Shapiro, this is a mistake: "Kim's reasoning rests on a confusion. In fact, the probabilistic equality above is not true, and in any event it is the wrong thing to consider if we want to know whether M causes P*. On the

standard definition of conditional probability, the right hand side of the equation is undefined. [...] If P is the supervenience base of M, then P cannot be present while M is absent" (2010: 600). So, as before, we have the claim that, because of the supervenience relation, P must bring M along. With this view on board, Shapiro proposes an empirical test for the causal efficacy of the mental.

According to Shapiro, we can construct an experiment to test whether the mental is causally efficacious (2010: 601). The test goes roughly like this. In testing whether *x* or *y* causes *z*, we normally hold one of *x* or *y* fixed, "wiggle" the other, and see whether *z* still occurs. Something similar can be done in the case of the mental. Because M and P are tied together by supervenience, it is impossible to hold P fixed and wiggle M, so we must do something else. M and P will have a physical cause further back, P_0, that is responsible for their occurrence. So, we can hold P_0 fixed, wiggle M (which will force P to wiggle as well), and see whether that affects the outcome of P*. In wiggling P or M, P* will be affected, so *both* P and M are causally efficacious. So the causal efficacy of the mind can be saved.

Shapiro claims that Kim is confused, but I believe Shapiro may be misunderstanding Kim's point. As I mentioned earlier, independence may not be what's important: distinctness is. Kim is not making a claim about the probability of P* happening in our world if we divorce M from P *in our world*. This would be a bit silly, given that Kim himself acknowledges the supervenience relation. Kim's claim is metaphysical. A perfect physical duplicate of our world in which M is not present will still adhere to Causal Closure. Hence the claim that the mental is not causally efficacious in our world. Both worlds will have all of the same physical relations,

and the world without mind plays out just like ours. Shapiro is placing too much stock on his empirical test. Either M and P are identical, or they are not. If they are not identical, then the supervenience relation will be either weak, global, or some other variety weaker than logical and there will be worlds in which one exists without the other. If they can come apart and it turns out that P is the only one that is causally efficacious, then mentality is not causally efficacious in our world. In the case of global supervenience, we would need only a physical duplicate of our world in which P and P* are set up as before, but M is absent because of the displacement of a single electron four light years away from the action. Remember, we are considering a world that is a physical duplicate (or near enough) of ours. To claim that no such duplicate without mentality exists is to suggest identity strongly. Whether we want the identity claim or not, it is simply not built in to the concept of supervenience that the physical brings the mental along *in all worlds*. For Kim's exclusion argument to work, the only thing we need is the possibility of P without M, and that possibility holds unless we're prepared to make a very strong claim about the supervenience relation. Absent that claim, the mental is not efficacious, and the nonreductive panpsychist and physicalist are not truly offering physicalist theories. As such, Kim's causal exclusion argument against the nonreductive physicalist succeeds.

One minor point worth considering here. Perhaps the nonreductive physicalist can bite the bullet and accept rampant overdetermination.[25] In what we considered above, the No Overdetermination principle was accepted, and the objective was to prove that mental causation does not violate it. We can, alternatively, reject the principle. Suppose this is right:

[25] I say "bite the bullet" only because most nonreductive physicalists would strongly object to this response.

our world is subject to widespread, common overdetermination due to mind and matter. It should be possible to have worlds in which mental events cause physical events without themselves supervening on any physical events. For instance, if the mental weakly supervenes on the physical in our world, that leaves open worlds that are mental duplicates with physical gaps in causation. Perhaps in getting stabbed, I lack the requisite C-fiber complex or any adequate replacement, but nonetheless my pain causes me to yelp. We need at least one of two things. We'll either need an account of what makes it the case that this mental stuff gets to count as physical, given the lack of a physical supervenience base, or we'll need an interactionist account of how this causal interaction is even possible such that we may be able to save Descartes.

Have I established that the theory on offer is not a physicalist theory? Not decisively. I have been appealing to a logical supervenience relation, but there is an alternative type of supervenience that is often held to be different from logical supervenience. For instance, one may hold a metaphysical supervenience relation between mind and matter, such that, while the relation is not a logical identity relation, we still won't be able to find any worlds in which the two come apart. But we do not have an argument in favor of metaphysical supervenience. Indeed, we cannot have one. The most we can do is assert the position without defense. Perhaps that is too strong. We could certainly appeal to theoretical virtues such as simplicity, though if we go down that road, I believe identity will win out. Furthermore, maintaining that the relation is nonreductive supervenience threatens, as I've argued, to discount the view as being physicalist at all. As such, there are two claims I can make here: a strong and a weak

claim. The strong claim is that nonreductive panpsychism is not physicalist, and while I have not established this decisively, I believe I have given ample reason to doubt that it can be a truly physicalist theory. The weak claim is that nonreductive panpsychism is plagued by too many problems for us to take it on board. We need a theory that is definitely physicalist in order to close the explanatory gap. The weak claim is my official stance here.

Merely claiming that an identity holds between the protophenomenal and physical does not help us much. While an identity would certainly make the theory physicalist, it brings no comfort if it cannot be made sense of. Chapter 3 is dedicated in its entirety to making sense of what it would mean for the protophenomenal and physical to be identical. However, before we get to that, there is one more issue I must take a stance on. It is now time to consider how I should conceive of protophenomenal properties. How like or unlike our own phenomenal experiences are they? This is important, because on certain characterizations, the view will fail to be nonemergentist. To that I now turn my attention. Once finished, we will have a complete, precise characterization of a panpsychist view that can close the gap. I end the chapter by drawing out the promise that this view can get the job done.

Section E: Protophenomena

In an effort to provide a theory that avoids making mind emergent, I have opted for a micro panpsychist view that places protophenomenal properties at the fundamental level of reality. There are two ways this might work. The first, as we considered in the previous section, is to hold that protophenomenal properties and physical properties are different properties of

fundamental entities. The second is to hold that every protophenomenal property is identical with some physical property. I have rejected the first option on the grounds that it fails to provide a physicalist theory. This leaves us with the second option. The question I must concern myself with now is how best to conceive of these protophenomenal properties. This is our immediate task.

In Bennett's paper "Why I Am Not a Dualist," she concerns herself with the property dualist's attempts to close the explanatory gap. She believes that the property dualist is going to encounter an insurmountable problem in characterizing protophenomenal properties. That problem is that any way we might construe these protophenomenal properties will invariably give rise to the explanatory gap. Now, her target is specifically a nonphysicalist form of property dualism; she is not targeting the view that we are proposing. Nonetheless, what she has to say has serious, direct implications for how *we* ought to think about protophenomena. If her concerns hold water, that will spell trouble for both the property dualist and myself. Thus, it will be fruitful to work through what she has to say. Additionally, she provides a rough outline at the end of her paper for the kind of view I would need to defend if I wish to claim that protophenomenal and physical properties are identical. It will serve as a nice starting point for what's to come.

As Bennett describes it, the task of the property dualist is strikingly similar to my own. In an effort to explain consciousness, the property dualist posits phenomenal (or protophenomenal) properties as fundamental properties of reality in the hopes of providing a nonemergentist theory of mind. The property dualist believes that "there are some unfamiliar,

fundamental phenomenal or quasi-phenomenal properties out of which the familiar person-level ones are somehow built. There are common elements that combine and recombine in various ways to generate experience as we know it" (Bennett, 2021: 13). Through laws of combination, she suggests, we can explain how these simpler phenomenal properties come together to form our minds. If we can be successful in offering the rules of combination, then we can close the explanatory gap: we'll have a complete explanation, in principle, of how phenomenal experiences like ours come to be.

Unfortunately, the project is hopeless for the property dualist, Bennett believes. Before the property dualist can provide a picture of how the protophenomenal combines to form the phenomenal, we must have an understanding of what the protophenomenal is like. In answering the "crucial question" of "just how phenomenal these protophenomenal properties are supposed to be," the property dualist will have to deal with an unpleasant dilemma: "either a version of the hard problem rearises between the protophenomenal and phenomenal, or else a version of the hard problem rearises between the physical and the protophenomenal" (Bennett, 2021: 15). The hard problem of consciousness just is the explanatory gap, and Bennett thinks that whichever way one characterizes protophenomena, one will have to deal with the explanatory gap in one form or another. The problem she raises for the property dualist is one I will have to deal with as well. I, too, must say something about what protophenomena are like.

According to Bennett, there are only two ways we might imagine the protophenomenal properties to be like: they can either possess the marks of the mental or they

can lack the marks of the mental. Here are some of the marks of the mental that Bennett picks out: "First, there is something it is like to have them. Second, they are introspectible; we have a certain sort of privileged access to them. Third, that access is arguably incorrigible…" (2021: 15). Before we continue, it's worth noting that it needn't be the case that the protophenomenal is characterized as either possessing all three or lacking all three of the proposed marks. In fact, the only mark that I believe is important is the first: there is something it is like to *be* the thing that possesses protophenomenal character. Anyway, our options are simple. Either protophenomenal properties possess the marks of the mental or they lack them; there's either something it's like to be whatever possesses them, or there isn't. Whichever way we go, Bennett believes we face a gap.

Let's begin by considering the possibility that protophenomenal properties lack the marks of the mental. If this is the case, then "the explanatory gap has not been closed; it has just been shunted into the space between the protophenomenal and the phenomenal. The hard problem rearises there" (Bennett, 2021: 15). I believe that this is clearly correct. The objective behind positing the existence of protophenomenal properties in the first place is to explain consciousness like ours. We have already seen that this cannot be achieved by appealing to purely unminded, physical properties. But protophenomenal properties that lack the marks of the mental are no different from physical properties in their inability to offer us an explanation. There's nothing it's like to be them at all; there is no experience there. So, we'll face the same problems that the classical physicalist theories of mind face.

What if we say that protophenomenal properties do possess the marks? Bennett believes that the exact same problem arises if we go down this road. She asks that we entertain a possibility: "Let us, then, consider the claim that protophenomenal properties *are* introspectable, that carbon atoms have privileged access into their protophenomenal states, and that there is something it is like to be a carbon atom" (Bennett, 2021: 15). Before saying more, it's worth pointing out that this is a highly uncharitable way of presenting this option. As I mentioned earlier, we need not say that the protophenomenal possesses every conceivable mark of the mental. At the start of the chapter, I said that things like "thought" and "consciousness" are highly anthropocentric concepts, and claiming that the protophenomenal qualities of atoms are introspectable strongly suggests a capacity for thought that we need not posit. Suffice it to say that we are considering the possibility of the protophenomenal properties having a "what-it's-likeness." I take it that this is rich enough of an internal life as it is without having to grant something as robust as an ability to introspect to the humble electron. Now, if the protophenomenal possesses the mark of the mental, then, Bennett argues, a different explanatory gap arises: "If protophenomenal properties are so like phenomenal ones, well, then now we need a story about how the protophenomenal arises from the physical" (2021: 16).

The claim that a new gap arises between the protophenomenal and the physical given that the protophenomenal possesses the mark of the mental is false. Neither the property dualist nor I need accept the claim. Remember, according to both the property dualist and the view we're considering, the protophenomenal *does not arise* from the physical. It is

fundamental. Indeed, to claim that we "need a story" for how the protophenomenal arises from the physical is a curious claim, given what she says elsewhere. Throughout her paper, Bennett seems to have a lucid understanding of what the property dualist is claiming. In one instance, she claims that on the property dualist's view, "there are some unfamiliar, *fundamental* phenomenal or quasi-phenomenal properties out of which the familiar person-level ones are somehow built" (Bennett, 2021: 13, *emphasis mine*). On the second page, she contradicts the claim we're currently addressing, saying that "there are facts about phenomenal consciousness that *cannot be explained in purely physical terms*..." (Bennett, 2021: 2, *emphasis mine*). And, further on down she grants that "[t]he property dualist's claim is that the phenomenal properties, or at least protophenomenal properties, are among the *basic furniture of the world*" (Bennett, 2021: 2, *emphasis mine*). Thus, it's quite unusual that she would then proceed to claim that the property dualist faces a new gap in explanation between the physical and protophenomenal. No such explanation is forthcoming: the physical and protophenomenal properties are both taken as *basic* on this picture.

As an important aside: one thing we might note is that it is peculiar, on the property dualist's picture, that these nonphysical properties hang around with the physical properties as they do. Whether they supervene or not, it seems a curious feature of the world that the two always stick together if we don't have a story of the kind Bennett is requesting. Perhaps, even taken as fundamental, what Bennett is getting at is that the property dualist will need to say more about the relationship between the physical and protophenomenal to justify the presence of the protophenomenal. With this, I agree. However, such a concern will not prove

problematic for my view. I am offering to draw an identity between protophenomenal and physical properties. If successful, it will be as unsurprising that the two are always found together as it is that Clark Kent and Superman always seem to be in the same room to those privy to his secret.

Returning to our discussion about protophenomenal character, protophenomena must possess the mark of the mental if we are to avoid the explanatory gap. In order to explain the phenomenal experiences that we are familiar with, it must be the case that our protophenomenal building blocks, while quite different from our everyday experience, are not completely alien either. How should we think of them, then? There is more than one adequate way, and I don't want to commit fully to any particular plausible option. I believe we should not envision the possessors of protophenomenal qualities as having a complex inner life, where 'complex' means something like 'structured' or otherwise similar to the intelligibility of our own. The distinction between protophenomena and phenomena consists in a difference in degree, not in kind. Given that I am taking this commitment on board, I now wish to replace the prefix 'proto' with 'micro'. The term 'protophenomenal' and its ilk is normally understood to pick out something that lacks the marks of the mental. Since the experiences of fundamental entities are like something, I will refer to them as 'microphenomenal', 'microexperiential', and other such 'micro' terms. Now, what is the difference in degree? There are at least two ways we may conceive of this.

First, I offer what I take to be the more natural conception of microphenomenal character: it is very weak. While the following is not an argument, it can hopefully serve as an

illustrative example of how we should conceive of these properties. One's visual experience is quite sharp at the point of focus. As we drift away from the focal point to the periphery, it's clear that experience becomes hazier. It isn't that our periphery is blurry (that wouldn't really describe what peripheral experience is like), it just seems less intense and vivid. Imagine, then, that we can assign degrees of experience on a numerical spectrum from 0 to 10, where 10 is the most vivid experience attainable. The experience at the focal point may be a 5. However vivid we may believe our own experience to be, it would be highly anthropocentric and a tad arrogant to think that we have achieved the greatest vivacity of experience the universe has to offer, so I opt for the middle of the spectrum in our case. Perhaps at the midpoint between the edge of our visual field and the focal point of visual experience, we find that the vivacity of the experience is a 4.98. Maybe at the very edge of our field of view, where things are least detailed, we can call that a 4.95. I propose that on this system, the experience of the possessors of microphenomenal properties such as electrons may be attributed a 0.001. The difference in degree is quite substantial, and the inner life of an electron should be understood as being very basic. Let's call this microphenomenal characterization:

> *Weak Character*: Microphenomenal character blacks out at a dim 0.001; there is something it is like to be an electron, but it isn't much.

What would that be like? I haven't the foggiest idea. But, I feel certain that it would be like *something*. Maybe not much, but certainly something.

Second, microphenomenal character may swing in the opposite direction. Perhaps the more natural possibility of Weak Character is wrong—perhaps microphenomenal character is

more vivid than our own, maybe even maximally vivid. Where our phenomenal character is a 5 on the vivacity scale, microphenomenal character maxes out at a 10. I label this view:

> *Strong Character*: Microphenomenal character whites out at a vivid 10; there is something it is like to be an electron, and it is like everything.

The concept of "whiteout" employed above is not original to me. In Philip Goff's book, *Consciousness and Fundamental Reality* (2017), he cites an unpublished paper by Keith Turausky in which Turausky offers this interesting way of thinking about microphenomena.[26] Where we normally would think that as we get closer to the fundamental level of reality experience becomes dimmer, perhaps it actually swings the other way. Now, as mentioned above, we shouldn't think of this increase in experience as an increase in complexity, such that we can expect electrons to have meaningful, structured experiences. Rather, where experiential blackout is like almost nothing, experiential whiteout would be like everything at once (Turausky mentions, metaphorically, that it would be like hot and cold, soft and hard, etc. simultaneously). Such an experience would be utterly incomprehensible to the entity experiencing it. If electrons possess Strong Character, then as systems become more complex, experience is filtered out. This possibility is not new. Aldous Huxley, in his book, *The Doors of Perception*[27], considers that the brain, rather than being a consciousness generator, is a 'reducing valve' that takes in raw experience and then spits out "a measly trickle of the kind of consciousness which will help us to stay alive on the surface of this particular planet" (2011: 8). I don't see anything obviously wrong with this possibility.

---

[26] Keith Turausky is a graduate student at the University of Texas, and the paper in question is "Picturing Panpsychism: New Approaches to the Combination Problem."
[27] Originally published in 1952.

The above has been a positive account of our options for microphenomena, but there is something negative it is important to add. We must avoid the prejudice that whatever the microphenomenal character of simples turns out to be, it must be static. This is particularly so if electrons turn out to possess Weak Character. Perhaps the Weak Character of electrons truly is static, such that whatever it is like to be an electron is unchanging, but perhaps it isn't. There is nothing in our experiences that suggests that any kind of experience must be static. Maybe when electrons repel protons, their microphenomenal character is affected so as to mirror that interaction (a crude form of proto-sensation). Certainly we shouldn't believe that anything possessing Strong Character would be static, even when remaining perfectly still. My point is not that we ought to believe in dynamic character, but rather that we ought not be committed in either direction without external reasons for preferring one option over the other. If it turns out to be theoretically fruitful to buy into a static character, then we have positive reason to go that way; if the alternative holds, then dynamic character it is.

For the remainder of this dissertation, I will proceed as though Weak Character is what we will find at the fundamental level unless otherwise specified. I do not believe that we are likelier to find Weak Character, but it is an easier pill to swallow for most, and it'll be dialectically useful to proceed with this characterization. Thus, we now have a panpsychist view that makes the following claims. Microphenomenal properties are to be found ubiquitously at the fundamental level of reality. They bear the mark of the mental: there is something it is like to be an electron. Finally, these properties are identical with physical properties. If this is correct, then we have a theory that meets our necessary conditions for

closing the explanatory gap. The theory certainly takes qualia very seriously. It is a nonemergentist theory: experience does not ever emerge from nonexperience, and it turns out that experience is ubiquitous. Finally, the theory is physicalist. Or, at least, it promises to be.

Given that the theory meets these necessary conditions, we can now imagine what an explanation of consciousness would need to look like. For any given phenomenal experience, that experience can be explained by appeal to its physical-phenomenal components. These experiential components must combine in some way to form the phenomenal experiences of the kind we normally have. Furthermore, these experiences must reduce without remainder to their experiential components. Given this, the gap is closed. That this kind of experiential combination can occur is not clear, however. Indeed, this problem, which has long been known as the 'combination problem', has been a major obstacle for this kind of atomistic panpsychism for some time. William James, who greatly sympathized with the panpsychist project, believed the combination problem to be deeply troubling for micro panpsychist views like the one I've characterized (Skrbina, 2007). Goff believes the problem just to be the explanatory gap by a new name and to be insurmountable (2006). I disagree, and I will provide a way forward in the final chapter.

We must now turn our attention to the claim that the microphenomenal and physical are identical. This needs to be made sense of. In an earlier draft of Bennett's article (from 2005), she actually provides us with an excellent description of what we need to claim in what follows.

> The trick is to say that the protophenomenal properties themselves constitute or ground physical properties, and consequently that there can be no genuine question of

> how the protophenomenal arises from the physical. The idea is supposed to be that there is independent motivation for the view that physical properties and entities can be characterized only relationally, by their causal-dispositional roles. If such a view is correct, there is a pressing question about what intrinsic properties fill these causal-dispositional roles. One answer to this question is designed to also address the hard problem. If protophenomenal properties fill the causal-dispositional roles, we solve two problems at once.[28]

The broad view is this. There is a metaphysical problem concerning properties and dispositions. It is unclear what is supposed to serve as the causal basis for the dispositions of simple, fundamental entities. I will argue that microphenomena serve that role. Furthermore, the microphenomenal causal bases are identical with their physical manifestations. Clearly, there is much work ahead. If successful, it will be clear what it means to say that the microphenomenal and physical are identical. Furthermore, this will make it clear how the microphenomenal gets to be respectably physical. Finally, it provides an answer to an important question we raised in Chapter 1: What is phenomenal knowledge knowledge of? Our answer: It is knowledge of the intrinsic nature of matter.

[28] This was on page 18 of the earlier draft.

Chapter Three

Section A: Qualities and Dispositions

The thesis of this dissertation is that a certain brand of panpsychism is apt to close the explanatory gap. In the previous chapter, I provided a thorough characterization of the panpsychist theory that I believe can get the job done. Most of my efforts in the previous chapter were in ensuring that we had a theory that could resist being emergentist. This is because emergence, as I have argued, necessitates the existence of an explanatory gap. Given that the physicalist theories I considered in Chapter 1 all appeal to emergence, the explanatory gap plagues them all. Thus, if I am to have any hope of closing the gap, we must avoid emergence. To that end, part of my efforts involved embracing a micro panpsychist view. Merely allowing for the reduction of phenomenal states to microphenomenal states, however, is insufficient. The reduction lets us avoid emergence, but I am still at risk of creating a nonphysicalist theory, and thereby failing to meet the necessary conditions. For instance, the property-dualist-esque theory we considered in the previous chapter is unlikely to count as a genuinely physicalist theory. My proposed solution, then, is to identify the microphenomenal and the physical. The motivation here is straightforward. If the two turn out to be distinct, then it seems to follow almost by definition that the microphenomenal is not physical. Therefore, every microphenomenal property must be identical with some physical property or other, if the theory is to deliver on its promises.[29] But if it is to be at all interesting that this

[29] Given that the proposed theory is micro panpsychist, all physical properties will also be microphenomenal.

theory can close the explanatory gap, it must be the case that the theory is plausible in its own right. Thus, an explanation is demanded of me: what does it mean to say that the microphenomenal and the physical are identical? They seem, at the face of it, to be quite obviously distinct. It is this question that I wish to address in this chapter.

We are now in a position to phrase the proposal in more technical terms. When I claim that the microphenomenal and physical are identical, there are two claims that I am making: (1) Microphenomenal properties serve as the *causal bases* of physical dispositions; (2) All causal bases are identical with their dispositions. I will treat the second claim first. In what follows, I will provide a thorough defense of the plausibility of (2). I will then focus my efforts on establishing the plausibility of (1). The reasons for accepting (1), I admit, are less clear cut than those for accepting (2). It will be impossible to provide a decisive argument in favor of believing that the microphenomenal is the causal basis of physical dispositions. Indeed, it is impossible to provide a decisive argument for any characterization of the intrinsic nature of matter, as it is by necessity out of our reach. That this is so is something I will defend. Nonetheless, we will go through a number of reasons for believing that microphenomena are the likeliest candidates to serve as the causal bases of physical dispositions.

Let us begin, then, with the claim that causal bases are identical with their dispositions. This will require a detour into the metaphysics of properties. We'll start by considering an intuitive example that will allow us to have a firm grasp on the terminology we'll be employing—terms such as 'causal basis' and 'disposition'.

Consider a simple glass on a table. This glass has a number of properties. For instance, the glass is clear, it has a particular shape, it possesses a particular microstructure, it is situated on top of the table, it is three feet away from this chair, it is fragile, and so on. Some of these properties are clearly properties *of* the glass. Some of them are properties that hold *between* the glass and some other entity. That the glass is clear, for instance, is intuitively a property that is solely of the glass. That the glass is situated on top of the table, on the other hand, is a relation that holds between the glass and the table. In what follows, we will be concerned only with the properties that can be said to be properly *of* the glass. I will put relational properties aside. Note: a property is relational in this sense only if we can change the property without thereby changing the object at issue. We could, for instance, place the glass under the table (or place the table over the glass) without changing any of the properties of the glass itself. The properties we'll be dealing with are popularly referred to as 'sparse' properties, which are meant to be opposed to 'abundant' or 'mere Cambridge' properties.[30]

Now, on some views, a property is said to be anything that can be predicated of an object. This, however, will not do. Many things can be predicated of objects that are intuitively not properties of the object in question. Utilizing the phrase 'such that', for instance, allows us to predicate anything we like at all of an object (e.g., the glass is such that I am six feet tall). Furthermore, we can predicate of objects things like "clear-and-fragile" that, in predicate form, are a single predicate, but which clearly refer to two individual properties of the object in question. Thus, we shall be very cautious with our use of the terms 'property' and 'predicate'

[30] The use of 'sparse' and 'abundant' originates with Lewis. See Lewis (1983); Lewis (1986).

in our discussion. When I use the term 'property', I am specifically speaking of what we normally take to be actual, genuine properties of objects (even if it turns out that we are at times mistaken). When I use the term 'predicate', I will be speaking solely of a linguistic entity.

Let us return to the glass and its properties. Some of the properties I listed are qualities of the glass, whereas others are dispositions. For instance, the glass' microstructure is a quality of the glass. That the glass is fragile, however, is a disposition. Qualities are widely understood to be intrinsic properties of objects. Dispositions, on the other hand, are normally understood to be powers of the object. What is the relationship between the two? Well, what does it mean to say that the glass is fragile? To say that the glass is fragile is normally taken to be an expression of what would happen to the glass were certain conditions to hold. If the glass were to be struck, it would shatter. Notice that when expressing a disposition, we make an appeal to other entities. After all, the glass would need to be struck by something, be it a bat or the floor. However, this does not turn the disposition into a mere relation. In order to change the fragility of the glass, it would not suffice to change objects elsewhere: something about the glass itself would need to change. Notice further that the glass' fragility is in no way dependent upon the existence of other objects. A glass in a universe with the same laws as ours but where it is the only thing in existence is a glass that will never *manifest* its disposition, but it will nonetheless still possess it—it will still *be* fragile. But why is the glass fragile; what makes this the case? The answer is that it possesses a certain composition that is causally responsible for its shattering when struck. Thus, the *causal basis* of the glass is said to be its microstructure. Unlike the disposition, which manifests only when certain conditions hold, the microstructure

of the glass is supposed to be a *categorical* property. Categorical properties are always manifest, so long as the object is around. Furthermore, these categorical properties are the ones that tell the causal story of the manifestation of dispositions. Thus, the relationship between dispositions and causal bases is said to be one of dependence: the disposition is dependent upon its causal basis.

So, objects appear have two types of properties: qualities and dispositions. This terminology is largely in line with John Heil's terminology; I follow it closely here. Dispositions are taken to depend on some intrinsic qualities of the object in question. The fragility of this glass is somehow dependent on its microstructure. And so it's supposed to go for everything. If, as Simon Blackburn puts it, the causal explanations offered by the causal bases of dispositions "illustrate the doctrine" of dispositional dependence on qualities, then it's true that "the clock tells me the time *because* there is such-and-such arrangement of little bits inside it; Sandy barks *because* her vocal chords vibrate; the light glows *because* electrons whizz around in its filament," and so on (1990: 62, *emphasis mine*). This is, in loose terms, the standard view. The question now is: just how do dispositions relate to their qualities? As just mentioned, it seems that dispositions are somehow *dependent* on their categorical causal bases. Those causal bases serve as the causal explanation for the dispositions of objects. But more remains to be said. It is time we looked directly at the canonical view on the subject as defended by Elizabeth Prior, Robert Pargetter, and Frank Jackson. They claim that qualities and dispositions are distinct, and that dispositions themselves are causally inert. I will work through their arguments in an effort to lay out the most broadly accepted view of qualities and dispositions,

and I will pick out some of its odd consequences. I will then argue against this standard view. Specifically, I will argue that qualities and dispositions are identical, not distinct. From there, we will consider a worrying objection that targets the existence of qualitative properties before we get back to our discussion of microphenomena. For now, let us turn our attention to what Prior, Pargetter, and Jackson have to say.

Section B: The Canonical View

Prior, Pargetter, and Jackson (henceforth "PPJ") arguably set the canonical view on the relationship between dispositions and their causal bases in their seminal paper "Three Theses about Dispositions." They argue in favor of three theses (1982).

(1) The Causal Thesis: All dispositions have causal bases.
(2) The Distinctness Thesis: Causal bases are distinct from the dispositions they ground.
(3) The Impotence Thesis: Dispositions are causally inert.

We shall take a brief look at how they defend each of these. They begin by taking it as obvious that there must be a *reason* that the glass is fragile. They say of the fact that the glass is fragile: "This is not a miracle. There is, that is, a reason why the glass is fragile. This reason involves a causally relevant property (or property complex) of the glass, which we will call the *causal basis* of the disposition" (PPJ, 1982: 251). Now, apart from taking it as fairly obvious that all dispositions must have causal bases, they offer a defense that seems to hang largely on the truth of determinism. Whether their argument is successful is something we needn't touch upon here. That dispositions must have causal bases is largely uncontroversial and something that I

am happy to grant. The bulk of their efforts are focused on establishing the truth of the Distinctness Thesis and the Impotence Thesis.

Either the fragility of the glass is distinct from its causal basis, or it is identical with it. PPJ argue that we cannot identify dispositions with their causal bases, as doing so results in contradictions. Their reasoning is as follows. The causal basis of the fragility of this glass is some microstructural property of the glass. However, this porcelain vase is also fragile, and its fragility finds its causal basis in a very different microstructural property. If we were to identify dispositions with their causal bases, then we would end up identifying the fragility of the glass with its microstructural properties and the fragility of the vase with its particular microstructural properties. This would have the absurd result that the nonidentical microstructural properties of the glass and vase are identical. This cannot happen. According to PPJ, we "cannot say both that being fragile = having molecular bonding α, and that being fragile = having crystalline structure β; because by transitivity we would be led to the manifestly false conclusion that having molecular bonding α = having crystalline structure β" (1982: 253). Thus, it must be the case that dispositions, while importantly related to their causal bases, are not identical with them.

PPJ provide a second argument in defense of the Distinctness Thesis, though it is admittedly much weaker. Roughly, they argue that dispositions obviously have their causal bases contingently. This glass could have failed to be fragile in another world. However, to draw an identity is to make the fragility of the glass a necessary truth, which is an unacceptable consequence. This may sound familiar: it's an analogue of the modal argument used against

the identity theory. However, in this case, it's far from obvious that we could hold the causal basis fixed and rid ourselves of the fragility. I'll say more on this in the following section. In any case, the considerations above are supposed to lead us to accept that the fragility and microstructure of the glass are two distinct properties of it.

Consider now their defense of the Impotence Thesis. According to PPJ, the causal basis in conjunction with the relevant antecedent conditions of the disposition is sufficient to explain the manifestation of that disposition. Granting the disposition itself a causal role threatens rampant overdetermination. "This causal basis is a sufficient causal explanation of the breaking *as far as the properties of the object are concerned*. But then there is nothing left for any other properties of the object to do. By the Distinctness Thesis the disposition is one of these *other* properties, ergo the disposition does nothing" (PPJ, 1982: 255). Thus, the disposition itself must be causally inert. Consider a concrete example. We know that the glass is fragile, and we claim that the glass' fragility is grounded by its microstructure. When a hard object strikes the glass at adequate speed (meeting the antecedent conditions for fragility), the causal story is going to be something about breakage occurring between molecular bonds. At no point will the fragility itself play a causal role; indeed, there is no causal role for it to play.

What does this make the disposition out to be? PPJ aren't strongly committed to any one way of thinking of dispositions as properties, so long as their Impotence Thesis goes through (1982: 256). There aren't very many ways of conceiving of dispositions consistent with the Impotence Thesis. Indeed, I can think of only two. One way is to eliminate the

disposition altogether.[31] The other is to turn the fragility into a higher-order property of the causal basis, which they also entertain. Just what is this higher-order property? Well, what is certainly important to any characterization of what dispositions themselves are is that they make some counterfactual true of the relevant object. Such a counterfactual would be something along the lines of "were the glass struck, it would shatter," though it would need significant modification to avoid finkish cases.[32] On PPJ's picture, we can conceive of the possession of a disposition being nothing more than having a certain counterfactual hold true of the object (in virtue of holding true of the relevant causal basis). It is a property of crystalline structure β that, when struck, it results in the glass breaking apart as we expect fragile things to do. If this is right, then all it takes for an object to have a disposition is for a sufficiently robust counterfactual to hold true of that object. While these may be the only options open to PPJ, there is another way we may conceive of dispositions—as playing the causal role PPJ deny them. We will consider this alternative in a later section.

In what follows, I accept the Causal Thesis. As such, I won't say anything about it. Before I present any arguments against PPJ, it will be worthwhile to take a moment to notice a few peculiarities in their arguments. They determine that an identity cannot be drawn between a disposition and its causal basis. We'll evaluate whether this is true shortly, but what's important to notice here is that, were the two to be identified, it would be clear which property of the glass the fragility is. Without the identity, we're left looking at the glass without knowing

[31] They seem comfortable enough with this possibility, as they express in the very last sentence of their article. In the argument that follows a little later on against PPJ, the first step is to establish precisely that the dispositions they envision can't be said to exist. Granting that they don't exist allows us to skip a step, so nothing hangs on this.
[32] See Lewis (1997).

where we might find its fragility. Now, the immediate response to this may be "the fragility just is the glass' power to break." Certainly, this would get us toward a satisfying picture, but their third thesis turns the fragility of the glass into a ghost. After all, the glass' power seems to be entirely contained within its fragility's causal basis. The fragility itself is playing no role. If we wish to say that the fragility is simply that the counterfactual "were the glass to be struck, it would shatter" holds true, then it seems that we have transformed the fragility of the glass into a linguistic artifact. It would be preferable not to have causally inert properties of the glass that do nothing, especially when those properties go by the name 'powers'.

We are drawn to the Impotence Thesis by the Distinctness Thesis. PPJ are forced to accept that dispositions are causally inert because of two things. First, identifying dispositions with their causal bases apparently has contradictory results. Second, granting them both causal powers threatens overdetermination. I won't deny that this kind of overdetermination would be metaphysically problematic. However, they are mistaken that an identity cannot be drawn. Heil argues, quite convincingly, that the identity indeed holds. If the identity holds, then we need not accept the Impotence Thesis. Dispositions, which would be identical with their causal bases, would thereby count as powerful in their own right.

One final note on something peculiar about PPJ's theory. We might think that powerful categorical properties are something PPJ would want to deny the existence of. However, it turns out that they can't. Their own view grants the existence of powerful causal bases. If it didn't, then the glass would never shatter when struck. It is the glass' microstructure that, when hit, *does* what results in the shattering. By their own view, the dispositions *cannot do*

anything. But something still happens: the glass is struck and then breaks. If the disposition isn't responsible, then the causal basis is. This just means that the causal basis is powerful. This is what Heil claims: qualities are powerful. So, the difference between Heil's view and that of PPJ turns out not to be a major one. As far as I can see, there is but a single difference. In addition to the powerful qualities Heil defends, PPJ wish to add an additional, inert property that makes no difference to the world.

Let us now shift our attention to Heil's claims. That dispositions are causally inert is something we need not accept. Notice, I do not claim that we shall decisively prove that the view is false, but instead I will, in line with Heil, claim that the view he presents offers a better, more plausible alternative. Furthermore, the identity between dispositions and qualities is what the panpsychist theory we have been constructing will depend on. In order to grant dispositions back their powers, we need the identity, but PPJ's concerns loom overhead. So, how do we solve the problem?

One way is to claim that the identity is not between fragility and microstructures α and β, such that $\alpha = \beta$. Rather, the identity holds between fragility and some more specific shared property between the vase and the glass. Perhaps α and β both possess identical substructure γ, and it is substructure γ with which fragility is identical. In other words, it turns out that the glass and the vase do have the same causal basis for their fragility. While this is an option, I do not believe it can work, and my reasons will sound familiar. As in the case of the identity theory, it seems very unlikely that substructure γ will be possessed by all and only objects that are fragile. Furthermore, this would need to be the case for all dispositions. Is this possible? Yes.

Should we believe it to be the case? I believe not. It is far likelier that fragility is multiply realizable, such that many objects with different microstructures can count as fragile. Still, were this to work, we could have our identity. I believe, however, that there is a better response available.

PPJ made a mistake when claiming that fragility is a property possessed equally by the glass and the vase. While we can ascribe to both the predicate 'fragile', that predicate is picking out two distinct, albeit similar, properties. Why believe this? Our everyday language suggests that many different things can be fragile. However, we also accept that different fragile objects break in very different ways. Indeed, porcelain does not shatter in the same way that glass does. It would be overly demanding to have a distinct term for every unique instance of fragility, so we lump the different kinds of fragility under the same term. So, rather than believing that many different objects possess the selfsame property, "we should suppose that the predicate 'is fragile' is satisfied indifferently by objects possessing any of a *family* of properties…" (Heil, 2004: 234, *emphasis mine*). How do fragile objects come to belong to this family? It is not because they all possess the property of being fragile, since, as I've just mentioned, 'fragility' actually picks out a set of properties rather than being a property itself. Instead, 'fragility' is a predicate that can be satisfied by a number of properties. Perhaps all of these properties get to fall under this umbrella due to how they resemble. It may be difficult to make sense of precisely what is required of this resemblance relation, but I do not think that we need to worry about the details here. I find it satisfying to say that *we* believe they resemble, and we therefore employ the term. This is something that we already do with properties like color. We use the term 'red'

to refer to thousands of different shades indiscriminately.[33] When we identify the fragility of the glass with its microstructure, it does not thereby follow that the vase must have the same microstructure, nor that the vase's particular microstructure must be identical with the glass' distinct microstructure. Rather, the glass' fragility is identical with *its* particular microstructure, whereas the fragility of the vase is identical with its own particular qualities. It does not follow from this that the two objects have the same composition, any more than we should believe that the two objects are fragile in the same way, or red of the same shade. Dispositions amongst objects can be expected to be the same only in cases where the causal bases are exactly the same as well. This is suggestive. It gives us reason to believe that an identity is plausible.

As of right now, all I have said is that drawing an identity does not run into the problems PPJ raise. But this isn't enough. We must now consider positive reasons to believe that the identity actually holds, and why we should believe that the view offers us intuitive, good results.

Section C: Identity

We shouldn't posit dispositional properties just to keep them out of the causal picture. Doing so is, at a minimum, bizarre. In describing the peculiarity of such a move, Heil says: "In an effort to make sense of causal powers—dispositionality—Jackson and his colleagues posit

[33] We also could have cut up the color spectrum differently in any number of arbitrary ways. Nonetheless, those conventions would establish which shades get to fall under which color terms. What's important here is that the true property is the shade, not the term.

dispositions as higher-level properties. Having introduced these properties, they then express amazement that anyone could imagine that such properties might *do* anything. This is the kind of maneuver that gives philosophers a bad name" (2005: 349-50). Apart from the fact that we seem to have posited a new type of property to ultimately leave it out of the picture, doing so, I submit, brings along the threat of making the theory nonphysicalist. We are positing properties that literally do not interact with the physical world. Why would we lend existence to something that does nothing? This is a sentiment we have already seen before, as it has been expressed by Kim (1989). Surely, we want our dispositions to *do* something.

To see how we may grant them their due powers, we should take a closer look at how we ought to conceive of dispositions. We'll begin with an intuitive understanding of qualities. A quality is meant to be an *intrinsic* feature of an object. Furthermore, qualities are meant to serve as the causal bases of dispositions. What this means is that the qualities of objects are the *reason* that dispositions manifest. They play a central part in the causal story. When we strike a vase, "[i]f the vase should shatter, [...] this is not, strictly speaking, because it is *fragile*, but because it possesses a certain lower-level *qualitative property*" (Heil, 2004: 233-4, *emphasis mine*). There are two things we should note about this. The first is that the relation between qualities and dispositions is typically seen as some sort of *in-virtue-of* relation. The qualities of the object are the properties in virtue of which the dispositions can manifest. However, this in-virtue-of relation is infamously difficult to cash out. The idea is that the intrinsic qualities give rise to the manifestation of the relevant dispositions, but the question of *why* one thing gives rise to another is rather opaque. Furthermore, without an answer to this question, an

illusion of contingency arises. The second thing to note is that, in what was said about about fragility, it certainly appears as though we are leaving the disposition out of the causal story. But, I claim, this is a linguistic issue. In order to see this clearly, let's think about what we really believe dispositions to be.

The first thing we should notice about dispositions is that they "are intrinsic properties of objects possessing them" (Heil, 2005: 344). Remember what we said earlier in the chapter: a dispositional property is not a relational property, even though when we talk about such properties, we make reference to other objects. But the fragility of the glass does not disappear even if the rest of the world does. The glass will remain fragile even if it is the only object in existence. This is because the fragility of the glass *is an intrinsic property* of it. Furthermore, the view that the causal bases of dispositions are inert is false. All of the intrinsic properties of objects, from their fragility to their redness, are *powerful* (Heil, 2005: 346). That this is the case has effectively already been granted by PPJ. They claim that the particular microstructure of the glass is the only property that does anything while the glass shatters. This is a power. And, indeed, *all* of the glass' intrinsic properties causally contribute to the behaviors of the glass. Finally, the dispositions of objects are *actual properties* of the object, not merely possible properties. "A ball disposed to roll, a glass disposed to break, a salt crystal disposed to dissolve in water each possess some actual feature in virtue of which it *would* roll, break, or dissolve. A disposition is actual. What need not be actual is the manifestation of a disposition" (Heil, 2005: 344). As already stated above, it is no less true of a glass in an empty world that it is

fragile. Just as it is no less true that a red glass is red even if drifting through the darkness of space.

The above considerations strongly suggest that an identity holds between qualities and dispositions. Fragility and the microstructure of the glass that serves as its causal basis are exactly the same property. Now, Heil seems to word his reason for drawing the identity in terms of parsimony, asking "[w]hy not dispense with the higher-level dispositional property altogether?" given that this higher-level property does nothing, and "the possession of [the qualitative property] would *itself* amount to the possession of a power" (2004: 233-4). But I believe that our considerations above reveal the error in not drawing the identity to follow from a linguistic confusion. The term 'fragile' may refer to one of two things. Either it refers to the manifestation of some behaviors of the object in question, or it refers to whatever is responsible for those behaviors. It cannot be the former. If it were, a glass would be fragile only while it shatters—that is, only while its disposition is manifest. But this is clearly not what we mean when we speak of the glass' fragility. We mean the same thing that we mean when we say that the glass is red. This redness is manifest only under ideal conditions. It must be within eyeshot and exposed to white light. But few would claim that it ceases to be red once we leave the room. The term 'fragile', then, finds a better referent in the microstructure of the glass. *That* the glass possesses such a microstructure *just is* to say that it is fragile. The microstructure itself is powerful; the disposition and the quality are the same: "A property's 'qualitativity' is strictly identical with its dispositionality, and these are strictly identical with the property itself" (Heil, 2004: 243). Note further that every intuitively intrinsic quality of the glass is

already believed to be powerful: "Being spherical is a manifest quality of a baseball." A ball's sphericity is the quality in virtue of which it appears spherical to us. Furthermore, "it is in virtue of being spherical that a baseball can, for instance, roll: sphericity is, it would seem, a power possessed by the ball" (Heil, 2004: 243).

Drawing this identity removes any possibility of contingency between dispositions and their causal bases. It also provides an understanding of the in-virtue-of relation between the two, as they turn out to be just the same thing. As Heil puts it, the identity "does not regard the dispositional and the qualitative as 'aspects', or 'sides', or higher-order properties of properties. A property's dispositionality and its qualitativity are, as Locke might have put it, the self-same property differently considered" (2004: 243-4). This is a good result. Indeed, to allow for contingency is to believe that the property of negative charge could attract other negatively charged particles in other worlds. But this is just to pick out a different property. Kripke's considerations concerning reference are particularly salient here. 'Negative charge' rigidly refers to that property which results in the manifestation of repulsion and attraction under certain conditions. To claim that it could have been otherwise is to claim that *x* could have failed to be *x*. But this is the worst kind of contradiction: an obvious one.[34]

But what of multiple realizability? Many things can be fragile, and identifying fragility with its various causal bases, as previously stated, threatens contradiction. But this isn't a real worry. That fragility is predicated of many objects is, according to Heil, a linguistic

[34] "Negative charge" is a rigid designator. What we are picking out is the property of repulsion and attraction to the relevant entities. To say that negative charge could have been otherwise is, at best, to say that we could have named a different property "negative charge," but that's not the issue at hand.

convenience: "We find it convenient to say that a teacup, a piece of slate, a pocket watch, and a gramophone record all possess the same disposition: being fragile," but to claim that they all possess the exact same disposition "seems unlikely: the objects shatter in different ways" (2005: 347). We call, for the sake of ease, many different properties by the same name. Things that tend to shatter or break easily are called 'fragile', though we have no reason to believe that they all possess the same exact property. Consider color. We can imagine a row of a thousand vases, each of a slightly different shade of white. Some are a little closer to cream, others more brilliantly white, still others with a sheen and some with a matte surface. For ease, we would certainly say "here we have a thousand white vases," and we would find the pedant annoying for contesting "no, you have a thousand vases, the first of which is white, the second cream, the third..." How impossible it would be to communicate if such a degree of detail were demanded of our predicative practices. And note that color is much easier to discern than the precise motions of shattering. We should expect language to abstract away from the very real differences.[35] So we don't have multiple realization of dispositions—not really. Identifying the disposition with its causal basis serves us no problems on this end.

Thus, it seems unproblematic to identify qualities and their respective powers. It just turns out that our linguistic practices give rise to the illusion of distinctness. But the reality is that that "which is qualitative is identical with that which is powerful, and both are identical with the unitary property itself" (Jacobs, 2011: 92).

---

[35] I am grateful to William Melanson and Joseph McCaffrey for the discussion that led to the creation of this example.

We must now consider a potential problem. The causal basis of fragility serves as the causal basis for other dispositions of the same object. This might seems like an issue. After all, the vase is not only apt to shatter when struck in virtue of its causal basis, it is also apt to roll and make a certain sound while doing so in virtue of the exact same causal basis. Can we say that both of these dispositions are the same disposition? I believe the answer is 'yes', but more needs to be said.

The above appears problematic only if we hold a certain false view of what dispositions are. We often describe dispositions in terms of how they manifest. So, that the glass is fragile tells me something about how the glass will behave when struck. But, the question is: struck by *what*? Perhaps a bat. When we speak of the fragility of the glass, we are not referring solely to the glass. We are talking about what happens when the *intrinsic character* of the glass comes in contact with the *intrinsic character* of the bat. This is not an issue of dispositionality, it is an issue of manifestation: what does it take for the glass' fragility to become manifest? Manifestations of dispositions take place between objects: the "manifestation of a disposition is a manifestation of reciprocal disposition partners" (Heil, 2005: 350). What this shows is that a disposition can have more than one kind of manifestation, given "different reciprocal disposition partners." Let's take an example. Remember that *all* of an object's intrinsic properties are dispositional—they are all powerful. So, this ball's sphericity is powerful. It grants it the power to roll downhill and make "a concave depression in a lump of clay…" (Heil, 2005: 350-1). When we talk about the ball's disposition to roll, we are referring to its sphericity. Just the same, when we talk about its disposition to make concave depressions in clay, it is the

same sphericity to which we are referring. Both dispositions are the exact same property; the difference is only in which reciprocal partners we are referring to. When talking about the ball's ability to roll, we are making reference to how the sphericity of the ball interacts with the surface properties of the ramp. When talking about the ball's ability to make impressions in clay, we are talking about how the ball's sphericity interacts with the compositional properties of the clay. Thus, the different powers are actually just talk of how *the ball's sphericity* interacts *with other objects' intrinsic properties*. But it's all reference to intrinsic qualities of objects.

So, not only are the dispositions of the vase to make a certain sound when rolling and to be fragile the same disposition, that very same intrinsic quality is likely responsible for the color of the vase. Consider this. Color is just a disposition to interact with light in a certain way. The causal basis of that disposition—the intrinsic property of the vase that is actually at issue—is its microstructure. That very microstructure is the vase's fragility: the exact same microstructure, when interacting with a bat, shatters.[36] What's different are the behaviors, not the disposition. In other words, the manifestations of the disposition vary in accordance with the given dispositional partners. There aren't three different properties—the microstructure, fragility, and redness—there is only the microstructure, and its behaviors when paired with different entities. This variance in behaviors leads to a lot of predication on our part.

[36] Indeed, it makes sense to claim that the redness of the vase just is its microstructure. If we put the vase under a purely green light, its color looks different. But the vase didn't change color. The best way to make sense of this is to say that the vase's disposition to look red just is its microstructure. That microstructure is the vase's power to look red under white light and black under green light.

I have been talking a lot about microstructures, but this is perhaps an oversimplification that is allowing me to sneak too much in. Consider a thin, blue, glass ball.[37] This ball, in addition to being blue, is also fragile and capable of rolling down hills. If what I have been saying above is right, these different dispositions—reflecting blue light, being fragile, and rolling—are all actually just one microstructure. But that's not really true. There is a lot encompassed by the term 'microstructure', and it isn't all identical. For instance, the fragility of the ball isn't merely the microstructure of the ball, but more specifically the bonds of its composing molecules and how many of them there are (not very many, given that the ball is thin). Its blueness, however, seems to ignore those specific bonds and be determined instead by the way the molecules located at the surface of the object interact with photons. The fragility of the thin, blue, glass ball, then, can't simply be identical with the ball's microstructure, but rather with some aspect of that microstructure, such as the number of molecules, how they're arranged, and the strength of their bonds. The blueness, on the other hand, is identical with only some small subset of those molecules—those at its surface that get to interact with photons. How do we make sense of this?

Thus far, I have been talking in a somewhat metaphorical way. I have said of the thin, blue, glass ball that its fragility is identical with its causal basis: some microstructural property. Furthermore, that microstructure is also the power of color in the glass. However, as I have suggested before, this isn't exactly right. The microstructure of the glass is composed of a bunch of molecular bonds. Those bonds are further dispositions. The bonds themselves are

[37] Thanks to Earl Conee for discussion on this example.

made up of further dispositions—those of their composing atoms. The different dispositions at the macroscopic level turn out to be nothing more than the dispositions of the fundamental entities at the bottom. It is at the microphysical level that we will find the qualitative causal bases that can be cleanly identified with a multitude of dispositions. Our glass ball is blue and fragile. That blueness is reducible to the dispositions of some subset of its composing particles, not to its entire microstructure. Similarly, its fragility is reducible to some other subset of the glass ball's composing particles. Now, notice that those subsets can have overlapping members. The particles at the surface of the ball have molecular dispositions that might play into the breaking of the ball when it is struck.[38] Consider one such particle. Its qualitative character, let's say, gives rise to its disposition to bond in some way. Furthermore, that same qualitative character gives rise to its disposition to interact with photons as it does. This is where we will find an identity. It isn't that the manifestations are identical; they clearly are not. However, as I have already argued above, a disposition is not its manifestations. It is the qualitative property that is responsible for those manifestations. And in the case of this singular particle, its qualitative character results in one type of behavior when bonding and in another when interacting with photons. If we put enough of these together, we will get the macroscopic behaviors of the thin, blue, glass ball without having to claim that the identity is as crude as fragility = microstructure.

[38] If the surface is blue because it is painted, then perhaps they have no role to play at all in ensuring the ball is fragile. If, however, the surface is blue because of the glass itself, then those particles are likely to play into the causal story of the shattering.

There is another, related problem to consider. We will take a closer look at our humble glass. If we closely inspect it—if we take a close look at its microstructure—what we will find are not categorical qualities of the glass. Instead, we will find more dispositions. The microstructure is, after all, an arrangement of molecular bonds. But a 'bond' is a disposition to hold together between separate entities. Now, this isn't immediately problematic. It suggests only that perhaps fragility has an intrinsic property further down than molecular bonding that serves as its actual causal basis. I've certainly been taking this for granted. Yet, the further down we look, the more dispositions physics seems to find. And this is all it seems capable of finding. Upon closer inspection, "we find things like an electrical charge at a point, or rather varying over a region, but the magnitude of a field at a region is known only through its effect on other things in spatial relations to that region. [... Science] finds only dispositional properties all the way down" (Blackburn, 1990: 63). At the very bottom, we may find entities like electrons, which physics characterizes completely by their dispositional character. Negative charge is a disposition to attract and repel other things; mass is a resistance to acceleration; etc. And so physics goes all the way down. Thus, we may want to conclude that dispositions exhaust reality. The world, we might say, is purely dispositional. If so, intrinsic properties don't exist, which spells trouble for our view.

Can this work? Can the universe be purely dispositional? Well, there is nothing inside the concept of dispositionality that rules out its serving as a causal basis for other dispositions. Indeed, McKitrick believes that dispositions can and *do* serve as causal bases, arguing that a "causal basis for fragility might be a particular type of molecular bonding. Plausibly, to have a

particular type of molecular bonding is to have a dispositional property. [...] If a type of molecular bonding can serve as the basis of fragility, say, then there can be causal bases of dispositions that are themselves dispositions." I argue that this cannot work if dispositions are seen as we characterized them in Section B: as counterfactuals holding true of objects. Given this understanding of dispositions, there are a couple of ways that a purely dispositional world could be characterized. In what follows, I argue that neither is plausible. We can opt for a different understanding of dispositionality which will solve the problems I raise in the following section, though that view will, I claim, basically add up to Heil's view. The world must, I claim, possess intrinsic, categorical properties that are responsible for the manifestation of dispositions. It cannot be the case that we live in a world of pure powers.

In what comes next, we will take a look at the two ways one might defend the pure powers view. I will provide an argument against both. Additionally, I will argue, for independent reasons, that we cannot do away with intrinsic properties.

Section D: Pure Dispositionality

Consider our glass once more. If the world is purely one of dispositions, it means that all of the properties we once thought were categorical turn out to be dispositional: the world is made up of pure powers. Thus, not only is the microstructure of this glass dispositional, but so is the color, shape, and so on. This may have stricken us as unintuitive at one point, but we have seen that physics tells us that the properties that we normally take to be categorical really are dispositional. Color—a paradigmatic case of a seemingly nondispositional property—turns

out to be nothing more than a behavioral manifestation. Objects can interact with photons through reflection, refraction, and absorption. Thus, to say that the glass is, e.g., red, is to grant it a disposition to reflect, refract, and absorb light a certain way. What of shape? It, too, turns out to be dispositional on the pure powers view. The shape of this glass is nothing more than a manifestation of the dispositions of the glass' components to bind together in a specific structural arrangement. The shape is merely an amalgamation of the dispositions of smaller entities. Now, are shape and color *purely* dispositional properties? For the moment, so as to entertain the pure powers view, we will suppose that the answer is 'yes'. Later, I will argue that this view of the world is untenable. Anyway, what it means for an object to be purely dispositional is for all of its properties to turn out to be powers and not qualities. Fragility is a disposition to shatter, redness a disposition to cause a certain sensation in us, and so on.

If all objects are purely dispositional, then the world comprises pure powers. Every object exists solely to push and pull upon the other objects of the world. How can this work? In order to avoid an infinite regress, there are two ways I can think of. The first is to suppose the world to be a dispositional network, where all objects are related to one another by their dispositions in a sort of web. The second is to grant that some dispositions are bare, grounding themselves and other dispositions. Let us consider these in turn.

If one believes that every property is a disposition in a network, one must find a way to avoid vicious circularity. After all, it might seem problematic to claim that entity A has a disposition to bring about entity B, which in turn has the disposition to bring about entity A. The solution is akin to how the functionalist picks out mental states through Ramseification,

and David Lewis has defended this method in the case of properties as well, though he makes use of qualities.[39] Thus, it must be the case that the world can be such that all objects exist in a causal network of dispositionality. Richard Holton argues in favor of the logical tenability of this view, specifically concerning himself with the counterfactual account of dispositions (1999). What might this look like? As I've stated previously, dispositions are often picked out by certain counterfactuals. Many offer counterfactual analyses of dispositions, such that all there is to a disposition existing is that some sufficiently robust counterfactual holds true of an object, which is the view we're currently considering.[40] These counterfactuals concern the manifestation conditions of the disposition. Thus, to say that our glass is fragile is just to say "were the glass struck, it would break." This, of course, is an oversimplification of a much more complex, more adequate counterfactual. Struck by what? Perhaps a bat plays into the antecedent of the true counterfactual for the glass' fragility. That bat, then, would be characterized by further dispositions, and those dispositions would make appeal to further objects in the world (presumably coming back to the glass at some point). Given that the world is causally closed, every object would sooner or later figure into the network of counterfactuals. Abstracting away from the particulars, we can construct a universe with only four purely dispositional entities. This is what Holton does in his "Dispositions All the Way Round." These entities, represented as points (though not meant to actually be points, as that threatens qualitative character), are fully characterized by their dispositional relations to the other entities. There are no categorical facts about them. So, entity A has nothing more to its

[39] See Lewis (2009).
[40] See Lewis (1997).

character than an ability to affect and be affected by B and C, which in turn have nothing more to their character than to affect and be affected by the other entities in that world.

There are at least two major problems with this view. The first, articulated by Blackburn, is that truth effectively disappears from such a world. Remember, a disposition is characterized counterfactually. Counterfactuals are what *would* happen in certain circumstances. In other-worlds talk, to say that it is true of the glass that it is fragile is to say that there is a nearby world where it has been struck and has shattered. But this is problematic, because it relativizes truth in our world to truths about other worlds. He says that "[t]o conceive of *all* the truths about a world as dispositional, is to suppose that a world is entirely described by what is true at *neighbouring* worlds," but the dispositions in those worlds are subject to the same problem; "the result is that there is no truth anywhere" (Blackburn, 1990: 64). I believe that Blackburn is correct. A world of a network of pure powers is a world in which only counterfactuals hold true, but nothing just holds true of the world itself. There is a second problem to consider.

If all there is to the character of an object is dispositionality, then the world we live in turns out to be the void. Heil provides a helpful example to illustrate the worry.

> Imagine a row of dominos arranged so that when the first domino topples it topples the second domino, which topples the third, and so on. Now imagine that, all there is to the first domino is a power to topple the second domino, and all there is to the second domino is a power to be toppled and a power to topple the third domino, and so on. If all there is to a domino is a power to topple or be toppled by an adjacent domino, nothing happens: no domino topples because there is nothing—no thing—to topple. (Heil, 2004: 237).

To characterize the world as possessing solely nonqualitative objects—to say that all objects are dispositional only—is to create a world in which there is nothing. It is just the void pushing upon itself. Perhaps this is too quick. There aren't just properties in the world, there are also the substances that possess them.[41] Thus, these counterfactuals hold true of *something*: an object. Doesn't this make the dominoes in the example quoted above not void-like? Well, consider one particular domino as PPJ would have us characterize it. The domino, on that view, possesses two types of properties. The first type are the qualities of the domino. These qualities are there to provide the domino with intrinsic character. They are meant to be properties of the domino and nothing else. Then, there are the additional properties which aim to tell us something about the behaviors of the domino. One such property would concern itself with what happens to the intrinsic character of that domino when I bump into it with my finger. That property is the counterfactual concerning my finger and its toppling. If I may speak somewhat metaphorically, what the pure powers view we're considering here is proposing is that we strip away the qualitative character of the domino. Thus there goes its color, so that there's nothing to be seen toppling; its weight, so that there's no mass to hit the ground; its shape, so that there's no boundary to come into contact with other dominoes; and so on. I was speaking metaphorically, and perhaps that's the problem with this picture. The pure powers theorist isn't claiming that I've tossed out the qualitative character, but rather that no such intrinsic character exists. Nonetheless, there's still plenty of *extrinsic* character. The color *is there*, it's just dispositional. However, on the counterfactual account, these

[41] I use 'possess' here in as neutral a manner as possible. I wish to not weigh in on the debate concerning the relation between substances and properties here.

counterfactuals don't seem to hold true of anything at all. That the domino is an object that is disposed to fall upon being hit is true of a qualitatively empty substance. This is strange, and stranger still that this empty substance might ever interact with its brethren, who display equal paucity of intrinsic character. I find the peculiar character that such a network of dispositions possesses striking: namely, no intrinsic character at all. It is difficult to even illustrate the subject at hand, because such a world is difficult, if not impossible, to envision. Heil minces no words on the matter: "A weighty tradition, going back at least to Berkeley, has it that the notion of a world without qualities is incoherent: a wholly non-qualitative world is literally unthinkable" (2004: 224). Perhaps I lack a certain kind of imagination, but I admit that such a scenario does indeed prove unthinkable to me. To be forthright, and perhaps redundant, one needn't accept the counterfactual account of dispositions to hold a pure powers view. We will return to this.

In any case, we needn't accept the network view of pure powers. Jennifer McKitrick proposes the possibility of bare dispositions. How is this supposed to help? In at least one way: we don't need to have a web of dispositions, we can have ultimate grounds for dispositions. It just turns out that some dispositions, likely those at the bottom, are bare (I read 'bare' as 'brute'). To be clear, McKitrick is making the claim that *some* dispositions are this way, not all. According to McKitrick, "*A bare disposition is a disposition that has no distinct causal basis*, neither dispositional nor categorical. A disposition whose unique causal basis is itself would count as a bare disposition. If an object has a bare disposition, the object has no intrinsic

properties which are both distinct from the disposition and causally relevant to its manifestation" (2003: 354).

As an important aside, in what follows, I concern myself only with the possibility of a pure powers view that makes use of non-qualitative bare dispositions, where the dispositions are counterfactuals holding true of objects. This is not McKitrick's actual view. She is perfectly content to accept that bare dispositions can, in addition to being dispositional, also be qualitative. In talking about Mark Johnston's views on bare dispositionality, she says "Johnston's definition significantly differs from mine in that it rules out bare dispositions that are both intrinsic and causally relevant to their manifestations" (McKitrick, 2003: 355). Thus, I am not necessarily targeting her theory. Nonetheless, some of what I say will have direct bearing on the plausibility of her position, as we'll see.

Why believe in bare dispositions? McKitrick has at least two major reasons for believing that bare dispositionality is possible. First, she takes it as plausible that dispositions can serve as the causal bases of further dispositions, claiming, as cited previously, "[i]f a type of molecular bonding can serve as the basis of fragility, say, then there can be causal bases of dispositions that are themselves dispositions" (McKitrick, 2003: 353). Second, our best physics reveals only dispositions at the lowest level of reality. Thus, it may very well be that negative charge is its own causal basis. Phrased less controversially, I take it that she means something like "negative charge could plausibly be brute," though this may be just an approximation. One objection to this view is the one raised above for the network view. Regardless of whether the dispositions are bare or not, a world of pure powers is no different from the void. Thus, I won't recite the

same argument again. Instead, there is a different argument I forward that proves problematic for McKitrick's bare dispositions.

McKitrick's argument seems largely to depend on the plausibility of one disposition serving as the causal basis for another. Certainly, if this turns out to be impossible, then it cannot be the case that a disposition can serve as its own causal basis. I will now argue that what McKitrick is conceiving of as a dispositional causal basis is actually just the disposition itself.

Consider the glass' fragility. McKitrick claims that it is plausible that the causal basis for this fragility is the glass' microstructure. This is to say that some molecular bonds serve as the causal basis of the glass' fragility. Now, these molecular bonds are supposed to be further dispositions. As McKitrick grants, if the bonds are dispositional, this means that they must be importantly related to some manifestation conditions. The same goes for the fragility of the glass. Well, what are the manifestation conditions for the molecular bonds of the glass? It turns out that they are exactly the same manifestation conditions as those for the fragility of the glass. Notice, it may not be the case that a singular bond within the glass has the same manifestation conditions as the fragility of the glass itself, but that singular bond will also not serve as the causal basis for the glass' fragility. It must be the collection of bonds which constitute the glass that serve as the causal basis. However, the antecedent conditions and manifest behaviors of the collection of molecular bonds are literally identical with those of the fragility of the glass. What this strongly suggests is that the molecular bonding is not serving as the causal basis of the fragility; rather, the bonding just *is* the fragility of the glass. The two dispositions turn out to

be the same singular disposition. Consider a different way of wording the point. The phrases 'fragile' and 'possesses such-and-such molecular bonding' are two predicates for the same dispositional property. This is a bad consequence for McKitrick's view, especially given that she claims that her "arguments proceed on the assumption that the issues are metaphysical, not merely linguistic" (2003: 353). But, of course, what I say above turns things into a merely linguistic issue. This does not show that no dispositions can serve as causal bases, but we quickly see that we have no obvious examples of dispositions serving as causal bases for further dispositions. Redness, we claimed, was plausibly a disposition to occasion a certain color experience in us. Following McKitrick, we might want to claim that certain surface properties of the object serve as the causal basis. However, those surface properties, being dispositional themselves, just are the disposition of redness of the object. With no plausible examples of dispositions serving as causal bases, we have no reason to believe that dispositions can serve as their own causal bases. The result is intuitive. After all, the objective of a causal basis is to bring about the manifestation of the disposition. However, dispositions as understood by the pure powers view are precisely the kind of thing that cannot provide this. They just are pure powers, but there's nothing that can provide an explanation for those powers.

I've stated now a number of times that I have been concerning myself specifically with the counterfactual view of dispositions. The main reason for this is that it's a fairly common view, and it's certainly compatible with PPJ's three theses. However, as I've mentioned, there is another way we can conceive of dispositions, and that is as powerful properties in their own right. This possibility is inconsistent with PPJ, as it would be in direct violation of the

Impotence Thesis. Granting that dispositions are genuinely powerful properties of objects, however, turns the disagreement between the pure powers view and the identity claim from the previous section into a semantic dispute. Consider: what Heil does is claim that PPJ's theory eliminates dispositions and grants categorical properties powers. Those categorical properties are *intrinsic* properties of objects, and it is those very properties that are powerful. Well, the view that dispositions are genuinely powerful doesn't seem very different. If the dispositions themselves are powers of objects, then those dispositions, too, are *intrinsic* properties of objects that themselves are responsible for everything the object does. Furthermore, such a view would have those dispositions sticking around regardless of whether they are exhibiting their manifestations or not: the glass is fragile, fragility is a genuinely powerful disposition, and as of right now that disposition is doing nothing as the glass has yet to be struck. The difference between the two views seems to be nothing more than a disagreement over whether we should call these properties 'dispositions' or call them 'categorical'. Furthermore, as McKitrick is happy to grant, bare dispositions can be qualitative. I don't see what grounds we would have for claiming that some are and some aren't. That the electron is negatively charged picks out both a disposition and a quality, the two being the same thing. There are some details I am brushing over, and I'm sure that greater differences between the two views can be drawn out. However, what matters for our purposes is that this identity be plausible, and on both the pure powers view as just described and Heil's view this seems to work. This type of pure powers view, as I see it, taints the purity of the powers such as to warrant not calling it a 'pure powers' view in this discussion.

The considerations above are sufficient, I believe, the deny that the pure powers view should be accepted. There is an additional reason that we should shy away from it, and that is the existence of qualia. Whatever else qualia may be, they are qualitative. Even if qualia can be somehow made out to be dispositional, they are more obviously qualitative than they are dispositional. They are, after all, *qualia*; it's in the name. Blackburn says of qualia: "Categoricity in fact comes with the subjective view: there is nothing dispositional, to the subject, in the onset of a pain or a flash in the visual field. Such events come displayed to us as bare, monadic, changes in particular elements of experience. In this perspective a change in perceived colour is as categorical as a change in shape or a twinge of toothache…" (1990: 65). Now, I deny Blackburn's claim that there is *nothing* dispositional in having a pain, but we can leave this issue aside for the time being. If qualia are necessarily qualitative, then any view that denies the existence of qualitative properties will deny the existence of qualia. If nothing else, the pure powers view is unworkable for my purposes.

If the world cannot be purely dispositional, then the negative charge of the electron must be as it is for a reason. In other words, it must possess a categorical causal basis. Furthermore, my denial of the pure powers view tells us this much: the world has qualities and dispositions. The pure powers view is simply a denial of qualities, but we cannot deny them. Physics, I have said, finds only dispositions in the world, but this should not lead us to believe that the world is made purely of dispositions. As we have just seen, the pure powers view is unworkable. But, furthermore, we should be unsurprised that this is the view that physics gives us. Physics cannot tell us anything about the intrinsic character of the world: it can deliver only

dispositions. As Heil remarks, physics "is silent on an electron's qualities[, but] it would be a mistake to interpret silence as outright denial [...], physics' silence on qualities does not amount to an affirmation that there are no qualities" (2004: 244-5).

Indeed, we have seen that there must be qualities. A world without them is a void; we should expect to find qualities at the bottom. David Chalmers says that the presence of dispositions leads us to "expect some underlying intrinsic properties that ground the dispositions, characterizing the entities that stand in these relations" (2003: 36). Grover Maxwell believes that, far from remaining silent, physics suggests that such intrinsic character must exist, saying that physics remains silent because we "can refer to such physical events only with descriptions or with terms whose reference has been fixed by means of descriptions or by other *topic-neutral*, non-ostensive means," but this, rather than suggesting that the intrinsic natures of fundamental entities don't exist, is merely a display of our scientific limitations: "It is just that our references to physical events by means of *topic-neutral designators* is an explicit signal of our ignorance of their intrinsic nature—our ignorance as to *what* such physical entities *are*. It is a reminder that our knowledge of them is limited to their causal and other structural properties" (1979: 396). Maxwell claims that physics is specifically picking out qualities in the world by reference to their causal interactions; it cannot be that there's nothing there being picked out. The silence of physics on what those intrinsic properties are is a consequence of its limitations. The scope and aim of physics, as Bertrand Russell says, "consciously or unconsciously, has always been to discover what we may call the causal skeleton of the world" (1927: 391).

While the fragility of the glass ultimately collapses into the dispositions of fundamental particles, we nonetheless require that these particles have intrinsic qualities. Maxwell argues that "science *does* assert the *existence* of instances of a variety of intrinsic properties; moreover, it provides information about the various causal-structural roles that such instances play" (1979: 397). We know these fundamental entities must possess intrinsic qualities. The question now is: "what are the intrinsic properties of fundamental physical systems?" (Chalmers, 2003: 36). My proposal: microphenomena serve as those intrinsic qualities.

Placing microphenomena here provides us with a truly physicalist theory. Chalmers puts the point nicely, stating that the idea of such a view is "that current physics characterizes its underlying properties (such as mass and charge) in terms of abstract structures and relations, but it leaves open their intrinsic natures" (2003: 25). If we want a complete story of physics, then "a complete physical description of the world must also characterize the intrinsic properties that ground these structures and relations," and if we make microphenomena fill these intrinsic roles, "once such intrinsic properties are invoked, physics will go beyond structure and dynamics, in such a way that truths about consciousness may be entailed." So we can give a complete physical story of reality, with emphasis on the word '*physical*'. Such a view serves to "retain the *structure* of physical theory as it already exists; it simply supplements this structure with an intrinsic nature. And the view acknowledges a clear causal role for consciousness in the physical world: (proto)phenomenal properties serve as the ultimate categorical basis of all physical causation" (Chalmers, 2003: 37). I will argue for the plausibility of this view in the next section.

Let's summarize. We now have a view of properties that identifies qualities with their dispositions. On this view, the qualities of entities are powerful in their own right. The sphericity of the ball is powerful: it is that very sphericity that makes it the case that the ball rolls down hills and makes impressions in clay. Furthermore, the same disposition gives rise to different behaviors depending on its dispositional partners. The ball rolls when in contact with hills and makes impressions when in contact with clay. However, these dispositions bottom out in the fundamental ultimates of reality, perhaps electrons. These electrons cannot be purely dispositional, so they must have some qualitative character responsible for their behaviors. However, it isn't clear what that qualitative character is. Whatever it is, the character is identical with the electron's physical manifestations. Indeed, physics concerns itself with finding these dispositions; that is the job of physics. So it is unsurprising that physics finds only dispositions. Whatever the intrinsic character of an electron may be, given that its dispositions are physical and that those dispositions are identical with its intrinsic character, the intrinsic character, too, must be physical. My suggestion is that the intrinsic character is microphenomenal. Indeed, there is a hole in the universe that needs filling: fundamental entities *must have* intrinsic properties that serve as the causal bases of their dispositions. This theory fills this hole with microphenomena. If this works, then, I claim, the explanatory gap is no more. In what follows, I will spell out precisely what the view is, how it's supposed to work, and I show that, if true, the explanatory gap of Chapter 1 is closed.

Section E: Microphenomenal Bases

We now have a view that identifies causal bases with their dispositions. This, of course, takes place at the fundamental level of reality, as that is where we will find qualities. Our proposal, then, is that these qualities are microphenomenal. So, an electron has a microphenomenal causal basis for its physical dispositions, such as its negative charge. What are some reasons to believe this? Here are a few. Placing the microphenomenal as the causal basis of physical dispositions gets the job done. As emphasized in the previous section, there is a hole in the universe that needs filling. As stated earlier, Heil rightly points out that physics "is silent on an electron's qualities" (2004: 244-5). The fact is, physical dispositions need intrinsic qualities. Indeed, assuming that they are identical, to deny one is to deny the other. Crudely, if nothing else, placing the microphenomenal in this hole gets the job done. Furthermore, the view is plausible. The only intrinsic qualities with which we are familiar are those of our qualitative experience. How fitting, then, that the insides of electrons consist of the same type of stuff. Indeed, Blackburn strongly suggests that the only way to save the world from the terrors of the void is to see its dispositional order "as a kind of construct from the categorical point-instances of properties available to the subjective view—a kind of neutral monism" (1990: 65). It also seems that nothing else can really serve the role of being an intrinsic nature. The stuff that we normally take to be intrinsic, such as shape, turns out to be nothing more than dispositions and relations between smaller entities. What could possibly serve as the intrinsic nature of a

simple, fundamental entity that lacks parts?[42] The only thing that even seems feasible at this level would be some kind of microphenomena. Is that really right? We have no idea what microphenomena are like, so why should we believe that this is plausible? Well, I claim that we do have *some* idea. I've already argued that the difference between phenomena and microphenomena is one of degree, not kind. Remember that a difference in kind would be subject to Bennett's powerful objections.

Finally, we ourselves have an intrinsic nature. This point is more powerful than it appears at first. We are already familiar with one kind of intrinsic property: qualia. It seems dogmatic, given the familiarity we have with qualia, to insist, without evidence, that the world outside *must* be devoid of qualia. Indeed, Russell claims that we not only have no reason, but that physics can stand to say nothing against the possibility: "To assert that the material *must* be very different from percepts [qualia] is to assume that we know a great deal more than we do in fact know of the intrinsic character of physical events[, ...] nothing that we know of the physical world can be used to disprove the supposition" (1927: 263). He says elsewhere: "I conclude, then, that there is no good ground for excluding percepts from the physical world, but several strong reasons for including them" (1927: 384). Russell believes it likely that percepts are 'compresent' with physical events.[43] Not only is it theoretically useful to claim that

[42] One potential view we could take is that there are no entities that lack parts. This kind of view of the world—that it is a plenum—is advocated by Margaret Cavendish. It's worth noting that she subscribed to a brand of panpsychism.
[43] It's worth noting that Russell is operating on an event-first ontology, though that has no effect on the current point.

microphenomena serve as the causal bases of physical dispositions, it even promises a way forward in making sense of consciousness.

This view may appear novel to many, but it is far from it (Skrbina, 2007). Blackburn, as we saw above, entertains it. Heil says something suggestive of the view, claiming that philosophers' "suspicions of *qualia* stem, in some measure, from more general suspicions of qualities per se. But if, as I have urged, everything has qualities, if every property is qualitative, then it would be a bad idea to treat putative mental properties as dubious solely because they are qualitative" (2004: 251). Now, Heil isn't embracing the view I'm defending here. Rather, it's clear from what he says that qualia are qualities in the same sense that the qualities of everyday physical entities are qualities. What's suggestive here is that it would take but a small step to grant qualia to all from what's been said. Indeed, he goes on to say that his view entails that zombies "are impossible." Chalmers has put the view forward in a few places. In his *The Conscious Mind*, he categorizes the view articulated here as a Type-C' view (1996). He considers the same view a Type-F Monism in his "Consciousness and Its Place in Nature," saying: "Type-F monism is the view that consciousness is constituted by the intrinsic properties of fundamental physical entities: that is, by the categorical bases of fundamental physical dispositions" (2003: 36). Jacobs appears to be amenable to this kind of view. He says that "qualia that are constituents of mental states are mental. Qualia that are not constituents of mental states are physical qualia. While mental and physical qualities are different in many important ways, none of those differences entail a special ontological status for mental qualia" (2011: 91). Now, Jacobs believes that qualia are not invariably mental, but by 'mental' he seems

to have 'conscious' in mind, which, as we've said previously, are not the same thing. He would likely be happier to say that they're mental in the way we've been treating the term, which is just to say that it has a sort of what-it's-likeness. Indeed, this seems to follow from his claim that there's no ontological difference between the two. Finally, he grants qualia inherent powers. He claims that he is not advocating for any kind of panpsychism, but, for better or for worse, his view certainly counts as panpsychist. Maxwell believed in a form of the identity theory, but his version placed the identities at a much lower level. In this sense, we, too, are advocating for a form of the identity theory, but the types are at the level of electrons and microphenomena, not brains and beliefs. Finally, Russell was partial to his neutral monism, the view being precisely that mentality, or whatever can give rise to it, is ubiquitous.

Here is one final reason to find the view plausible. Chalmers characterizes the view defended here as one that solves two problems at once: 1) where mental properties fit into the world, and 2) how dispositions relate to their causal bases. The solution runs like this:

> Perhaps the intrinsic properties of the physical world are themselves phenomenal properties. Or perhaps the intrinsic properties of the physical world are not phenomenal properties, but nevertheless constitute phenomenal properties: that is, perhaps they are protophenomenal properties. If so, then consciousness and physical reality are deeply intertwined. (Chalmers, 2003: 130)

The idea is that there's this one problem in the philosophy of mind, where we need to find how mind and matter are related, and then there's this other problem about how dispositions and qualities are related. I now wish to make a brief argument for believing that these two problems can be solved with one proposal. The debate over qualities and dispositions is a debate over the

intrinsic qualities of matter and its behaviors.[44] We see matter behave certain ways, and we believe those behaviors to be due to some intrinsic nature. We then have two questions: what is the intrinsic nature of matter, and what is the relation between that nature and its dispositions? But these are the very questions we ask in the mind-body debate. Bodies behave, and we want to know how those behaviors relate to the mind (our intrinsic natures). The only difference between these debates is that, in the mind-body case, we already have an answer to the intrinsic nature question: it's qualia. In the quality-disposition debate, we now have an answer to the other question: the relation between qualities and dispositions is identity. Finally, there is no reason to believe these to be two genuinely different metaphysical domains, though they are two different subjects. After all, however special humans may be, we are still material objects. The answers we provide in one area should have an effect on the other. Our brains are *material*, and so they must, just like everything else, have an intrinsic nature and dispositional character. That microphenomena serve as the intrinsic nature of matter and are to be identified with dispositions reveals that we are dealing with one question in both areas: what is the intrinsic nature of matter? In the case of all matter, including brains, it's qualia. As an aside, we already intuitively believe that our behaviors occur because of our minds. Perhaps this possibility provides a way forward: as Russell says, perhaps "the electron jumps when it likes..." (1927: 393). Maybe the behavior of electrons can be explained by the proto-desires of the electron. I take no stance on this issue, I merely suggest the possibility.

[44] In fact, the pure powers view can be seen as being a behaviorist theory of properties.

Is our theory truly physicalist? Yes. As Chalmers puts it, if "one holds that physical terms refer not to dispositional properties but the underlying intrinsic properties, then the protophenomenal properties can be seen as physical properties, thus preserving a sort of materialism" (2003: 130). The fact is, if the identity works, and we have no reason to believe it doesn't, then our theory is as physicalist as a theory can get. Indeed, I would say it is more plausibly physicalist than the theories we considered in Chapter 1. Is the view true? Maybe! I do find it quite promising. However, what I need for this dissertation isn't to establish that the theory is true. There are other competing theories that may very well also count as physicalist panpsychist theories that could get the job done. As such, I remain neutral on whether this particular theory is true. What we need is just for this theory to be plausible and meet the necessary conditions.

One perk of this theory is that it tells us precisely what phenomenal knowledge is knowledge of: it is knowledge of the intrinsic nature of matter. This explains why, as Conee argued, we can come to know phenomenal properties only through acquaintance. If phenomenal properties are our intrinsic natures, then of course facts could never deliver them to us—only direct acquaintance with them will suffice. When Mary first saw a ripe tomato, what she learned was something about herself: she learned something about the intrinsic character of herself—what it is like to *be* a certain way.

Let us now return to the explanatory gap. In the physicalist paradigm, no amount of physical information seems apt to settle the phenomenal. Thus, there is a gap in explanation from the purely physical to the purely phenomenal. What we need is a theory that can rightly

be called physicalist, that takes qualia seriously, and that avoids emergence. That our theory needs to be physicalist and take qualia seriously are straightforward requirements. That the theory needs to be nonemergentist is needed precisely because emergence, by necessity, eludes explanation. A theory that meets these requirements should offer us a way forward, and our theory does just that.

Before saying more, consider what I have done. I have identified microphenomenal qualities with physical dispositions. If this is right, then it's no surprise that the physicalist theories from Chapter 1 all failed. They operated under a faulty assumption: that the phenomenal was to be explained in terms of the physical. But we could never explain Superman in terms of Clark Kent. Phosphorus cannot be built up out of Hesperus. The only explanation that could possibly prove satisfying here is identity, and if we presuppose the rejection of that identity, then progress is impossible. By identifying microphenomena with physical dispositions, the way forward has opened up. The intrinsic character of an electron is not built up out of its dispositions: it is the intrinsic character of those dispositions. With this on hand, we are prepared to close the gap.

Our theory meets all three necessary conditions. I contend that the three together are sufficient to transform the hard problem of consciousness into another easy problem of consciousness, remembering that 'easy' here is relative. Perhaps better terms would be the 'impossible problem' and the 'Herculean problem'.

There are a few things I need to do to close the gap. According to Chalmers, a solution "to the hard problem would involve an account of the relation between physical processes and

consciousness, explaining on the basis of natural principles how and why it is that physical processes are associated with states of experience" (2003: 104). Now, my theory provides a direct answer to how the physical is related to the microphenomenal, so this isn't a perfect characterization of what I need to do to close the gap. What I mainly need to do is provide a reductive explanation. Chalmers says that a reductive explanation requires that consciousness be explained "wholly on the basis of physical principles that do not themselves make any appeal to consciousness" (2003: 104). This, however, isn't exactly right. In a footnote on the same page, he makes a more accurate assertion: "Reductive explanation requires only that a high-level phenomena can be explained wholly in terms of low-level phenomena." So, our explanation should be reductive.

Furthermore, our explanation needs to be combinatorial. This is for a few reasons. First, conscious states are complex: "Conscious states have structure: there is both internal structure within a single complex conscious state, and there are patterns of similarities and differences between conscious states. But this structure is a distinctively *phenomenal* structure..." (Chalmers, 2003: 122). In order to explain this complexity, it'll need to be clear that some mechanism of combination is possible. Heil and C. B. Martin say of the properties of complex objects that they "are wholly constituted by simpler component properties with all their qualitative and dispositional aspects for an infinity of reciprocal disposition partners, for an infinity of mutual manifestations within the limits of what they are *not* for and even what they prohibit" (1998: 290). What this suggests is that the same kind of story should be available

for the phenomenal. The microphenomenal will need to combine in such a way as to give rise to mind with all of its complexities.

The theory we have on hand is, it seems, prepared to meet these requirements. Phenomenal states are reducible to microphenomena. Furthermore, microphenomena, being the cores of fundamental entities, combine to form complex consciousness. There is something about the way that the raw material of our brain interacts that allows the microphenomenal qualities to add up to full-fledged conscious experience. Now, as Chalmers says, one might "object that we do not have any conception of what protophenomenal properties might be like, or of how they could constitute phenomenal properties. This is true, but one could suggest that this [is] merely a product of our ignorance" (2003: 132). This is not a problem. The point need only be that this is possible. We, similarly, don't quite know what it means for a quark to be 'strange'. Nonetheless, we can posit its existence and use it in the explanation of complex physical interactions. We can do the same with complex phenomenal states. That this is possible is enough.

Notice that the explanatory gap from Chapter 1 posed such an insurmountable barrier to progress that it wasn't even conceivable how we could move forward. With the theory we have on hand now, this is no longer the case. To cite Chalmers once more, in "the case of familiar physical properties, there were principled reasons (based on the character of physical concepts) for denying a constitutive connection to phenomenal properties. Here, there are no such principled reasons" (2003: 132). Unlike the physicalist theories from Chapter 1, this theory faces no in-principle objection. We have building blocks for phenomenology, and we

can imagine what it would be like to provide a story of combination. Even if we cannot characterize microphenomena directly, we may be able to characterize them theoretically, in the same way we have characterized the properties of quarks (Chalmers, 2003: 132). With a view like this, as Bennett says, we can almost *see* the microphenomena building up to our conscious experience (2005). We no longer need a miracle; we just need a story.

While the view here does do away with the original explanatory gap, there are those that believe that a new gap has formed—one particular to this view. According to philosophers like Goff, the combinatorial story we need to provide in order to explain how we get conscious states from microphenomena is impossible. Thus, a new explanatory gap rises up, and it's one we have to deal with. This new gap takes the name "the combination problem." Now, I believe that there is a mistake here right off the bat. The contention is that this explanatory gap is just as problematic as the original. That is not the case. The original gap was completely unbridgeable, and it was so problematic that it was inconceivable how the right story could even begin to be formulated. Against the combination problem, what we need to do is clear. There are a few versions of the combination problem, but against all of them the response required is precisely the same: we must show how combination is possible. We have every ingredient necessary to create conscious experience. The threat of the combination problem is that we will fail to find the right recipe, for one reason or another. Our response, then, is to show that the recipe is within reach. To do this, there are two things we will have to do.

The combination problems we will face in the next chapter will be of two kinds. The first kind will be those that make presuppositions that, if true, are problematic for our view. I

assert that we can safely deny those presuppositions without seriously modifying our panpsychist theory. The second kind rely on the assumption that the fusion of subjects is impossible. This is the more problematic version. To argue against it, I will provide reason to believe that subject fusion is, indeed, possible, and, further, plausible. I will not at any point be providing the precise story of combination. To do so would require a dissertation in its own right. Nonetheless, I will focus on making clear how that story can be provided. We must now turn our attention in full to the final problem facing my view: the combination problem.

Chapter Four

Section A: The Combination Problem

As presented in Chapter 1, the explanatory gap held between physical facts and phenomenal facts: there was no way of making sense of how one could get to the phenomenal from the purely physical. As we saw, standard physicalist theories ultimately made use of brute emergence in order to grant the mental a place in the world. But brute emergence, it turns out, necessitates the existence of a gap in explanation. After much work, my proposal has shaped up to be a particular kind of panpsychist theory. This theory places microphenomenal qualities at the fundamental level of reality. Indeed, the microphenomenal properties turn out to serve as the causal bases of physical dispositions.

My theory intuitively provides us with a way of closing the explanatory gap. The way we normally make sense of macroscopic properties is combinatorial. According to Sam Coleman, "higher-level properties demand to be understood as *configurational*: they are the mere product of the arrangement of lower-level bits and pieces *given the properties already in play*" (2015: 74). Indeed, the problem with standard physicalist theories of mind is precisely that no mental properties are to be found at the fundamental level, requiring that they emerge at the higher levels. My theory promises a combinatorial explanation. The properties of macrophenomena are to be made sense of by the combination of microphenomena. All that's left now is to find the rules of combination that lead these microphenomena to "build up" to

the more familiar macrophenomena of daily life. But this is precisely where my proposed solution to the explanatory gap runs into problems.

It is unclear precisely how the microphenomenal is supposed to *combine* to give rise to the phenomenal. This problem is commonly known as the 'combination problem', coined by William Seager, and Goff articulates the problem well.[45]

> On the most familiar versions, there are a huge number of micro-level (proto)subjects in your brain right now, each enjoying its own (proto)consciousness, which somehow come together to form, or to bring about, your mind and its consciousness. The essence of the combination problem is simply this: how on earth is that possible? We feel we have some kind of grip on how bricks forms [sic] a house or parts of a car engine make up an engine, but we are at a loss trying to make sense of lots of 'little' (proto)minds forming a 'big' mind. (Goff, 2017: 165).

The promise of my theory is that the microphenomenal can adequately combine to form the macroscopic experiences of our minds, but this raises an important question: how exactly is that supposed to happen? It turns out that providing a satisfying answer to this question is notoriously difficult, hence the *problem*. Thus, it is normally understood that a new explanatory gap emerges. Whereas there once was a gap between the physical and the phenomenal, there now exists a gap between the microphenomenal and the macrophenomenal. So, addressing the combination problem falls within our purview, and I must deal with it. And deal with it I shall. But first, there is something that must be made clear.

Contrary to popular sentiment, the combination problem is not like the original explanatory gap. We must remember that the original gap was a gap in explanation between supposedly nonexperiential physical properties and seemingly nonphysical experiential

---

[45] Seager coins the term on page 280 of his 1995 article, "Consciousness, Information, and Panpsychism."

properties. This explanatory gap was necessitated by the brute emergence of phenomenal qualities. Goff claims that the micro panpsychist must deal with the same problem, stating that "panpsychism is also committed to a kind of brute emergence which is arguably just as unintelligible as the emergence of the experiential from the non-experiential: the emergence of novel 'macroexperiential phenomena' from 'microexperiential phenomena'" (2006: 53).[46] This is a mistake. I grant that both the micro panpsychist and the classical physicalist must deal with some form of gap in explanation, but the gap of the classical physicalist is an unbridgeable chasm, whereas the micro panpsychist has a clear way forward. The task of the micro panpsychist is this: provide an explanation for how the micro-level phenomena can combine, merge, fuse, or whatever to form the macro-level phenomena. My claim is not that the task is easy, simply that it can be done. Where the original explanatory gap left us scratching our heads and confused to the point of accepting either magic (as Strawson deems brute emergence to amount to) or insuperable ignorance (McGinn's cognitive closure), we know precisely what kind of move the panpsychist needs to make. The classical physicalist was tasked with baking a cake without flour, and the response was that cakes arise from eggs and milk. The panpsychist has flour along with every other ingredient; the question now is just: what is the recipe? While the combination problem is pressing, difficult, and demands a solution, we must not set it on even footing with the original explanatory gap. The original gap is in a class of its own. Nonetheless, I will offer a way forward.

[46] It should be noted that this article was written prior to Goff's conversion to panpsychism. Nonetheless, he still holds that the combination problem is insurmountable, hence he buys into what he calls "intelligible emergence."

There are, I believe, a number of ways of solving the combination problem, and before I present it in explicit form, there is something worth remembering. Recall that in Chapter 2, I distinguished between two potential characterizations of microphenomena. There are at least two ways we might conceive of the microphenomenal lives of fundamental entities: they may exhibit either Weak Character or Strong Character. In offering theories of combination, I will ultimately provide two possible ways that we may make sense of what it means for microphenomena to combine, and I believe that each one of these two characterizations of microphenomena ends up being more amenable to one theory over another. For convenience, here's what I said about each kind of character in Chapter 2:

> *Weak Character*: Microphenomenal character blacks out at a dim 0.001; there is something it is like to be an electron, but it isn't much.
>
> *Strong Character*: Microphenomenal character whites out at a vivid 10; there is something it is like to be an electron, and it is like everything.

A quick note on the similarities and differences between these two types of character. The two are similar in an important respect: regardless of whether microphenomenal character blacks out or whites out, the microphenomenal experiences of fundamental entities are not structured, or meaningful, so to speak, like ours are. To be metaphorical, blackout is experiencing the TV while it's off, and whiteout is experiencing the TV display static: both are of equal use, and both are equally uninteresting. Where the two differ is just as important: if fundamental entities exhibit Weak Character, then combination results in an increase of experience; if they exhibit Strong Character, then combination filters experience out. It is this

difference that makes each character more suitable to one theory of combination over another, as will become clear later on. Let us now turn our attention to the combination problem.

The combination problem is not a single problem. There are actually a number of problems that are captured by the term. As such, it serves as a sort of umbrella for different questions concerning combination, and tackling them all would require too much space. Just as well, it is unnecessary to take them all on. Many combination problems do not give rise to the explanatory gap. For instance, Chalmers considers a combination problem concerning the structure of phenomenology (2016). According to certain micro panpsychist views, we should expect the structure of phenomenal and microphenomenal experience to be isomorphic with that of the physical structures in the brain, but we have reason to believe that can't be right. I don't know how to make sense of the supposed structural differences between the two, but this seems to be more of an "easy problem" of consciousness. A way forward may be difficult, but it is certainly within reach, and there doesn't appear to be any concerning gap in explanation.[47] Furthermore, the more pressing combination problems tend to be closely related, such that responding to one offers a way forward on many others. As such, I will focus on what I believe are the most concerning problems that cover the most ground. The problems we will consider are the following four: the zombie problem, the palette problem, the problem of the exclusivity of phenomenal states, and the subject combination problem. Briefly, the problems go like this.

The zombie problem is meant to be the panpsychist analogue of the classical philosophical zombies that plague standard physicalism. Part of the problem with the original

[47] Indeed, Chalmers proposes some ways we may move forward in that same paper (2016).

physicalist theories we considered was that zombies seemed perfectly conceivable. That we could conceive of zombies revealed a gap in explanation. The concern here is that the panpsychist, too, has to deal with a novel type of zombie. These zombies do have phenomenal experience, but their experiences are only the disparate micro-experiences. Their experiences fail to combine to form macro-experiences of the type we are familiar with. Any story of combination, it is claimed, is subject to the zombie problem.

The palette problem is, as the name suggests, a problem of how we are supposed to generate complex experience given an impoverished palette. Chalmers puts the problem clearly: "There is a vast array of macroqualities, including many different phenomenal colors, shapes, sounds, smells, and tastes. There is presumably only a limited palette of microqualities. [...] How can this limited palette of microqualities combine to yield the vast array of macroqualities?" (2016: 183). This is meant to be especially concerning given the different modalities of experience, such as sights and smells. How, then, could combination of a small palette of simple micro experiences give rise to the richness of macro experiences? The problem, it seems, is motivated by the following intuition: the experiences of things like us are unimaginably complex, whereas the experiences at the bottom just aren't. So, in order to account for this immense complexity, we will need far more basic ingredients than it will be plausible to believe there are.

The problem of the exclusivity of phenomenal states runs like so. Coleman believes that phenomenal states are characterized by what they include *and* what they exclude. Given that phenomenal states are exclusive, combination is rendered impossible. For instance, take a

blue-and-not-red experience and a red-and-not-blue experience. Now combine them. The resulting subject must somehow experience both, but this is a contradiction. Coleman provides perhaps the strongest phrasing, claiming that he shows decisively that combination cannot occur.

Finally, the subject combination problem claims that combination is impossible due to a special feature of subjects: "The problem is that conscious subjects seem to be in a certain sense irreducible: it doesn't seem that we can specify what it is for there to be a conscious subject in more fundamental terms" (Goff, 2017: 20). *We* are subjects, but micro panpsychism requires that our subjecthood reduce to our more basic experiential components. If subjects are irreducible, then combination is impossible, and the theory fails. This version of the combination problem is the deepest and most difficult to handle. It is, by far, the strongest version of the combination problem. Indeed, Coleman says that this version "is the real combination problem" (2013: 29). As such, it is worth it to draw out exactly what the problem is and why it seems so difficult.

Goff separates the subject combination problem into two: the subject-summing problem and the subject irreducibility problem. I believe that these are just two different ways of approaching the same issue. In any case, my proposed solution will ultimately handle both, so I'll treat them as the same. Goff defines 'subjecthood' as being anything that has an experience. He says: "I take it that subjecthood is a determinable of which each conscious state is a determinate. For example, to be pained is to be a subject in some specific way; to have an experience of orange is to be a subject in some other way" (2017: 178). On Goff's picture,

phenomenal properties are properties of subjects. As such, it is impossible to have an unexperienced experience.[48] Now, to be an experiencer just is to be a subject, and if all fundamental entities possess microphenomenal qualities, then those qualities are all experienced by subjects. I believe the most natural read here is that the possessor of the microphenomenal qualities is the subject, in the same way we consider ourselves to be the subjects of the phenomenal qualities we possess. However, this is not the only possibility.[49] As a reminder: these entities needn't possess capacities such as introspection; there being something it's like to be them is sufficient for my purposes. Now, it is clear that we, too, are things that experience: we, too, are subjects. Furthermore, we are *composed* of fundamental entities. This raises an important question: how is our subjecthood related to the subjecthood of our composing parts? This is the crux of the issue. The intuition is that subjecthood is irreducible, and this intuition coupled with the claim that our subjecthood is composed of micro-subjects generates the subject combination problem. Indeed, these taken together seem to make combination impossible.

A common move in dealing with combination problems is to acknowledge their difficulty and attempt to either lengthen the wick on the bomb or at least slow its rate of burning. My objective here is more ambitious. At least insofar as the combination problems we are addressing are concerned, my aim is to defuse the bomb. I aim to provide solutions. More

---

[48] That there are unexperienced phenomena is a route that is taken by certain micro panpsychists, especially Russellian monists. I won't be considering that view here. This makes the task at hand harder for me, so I take it that nothing is lost.

[49] For instance, one might believe that the phenomenal quality is its own subject, such that the subject is identical with the quality. Regardless of how we conceive of the relationship between subjects and phenomenal qualities, the same combination issues will arise and the same solutions will apply.

precisely, my objective is to show clear ways to move forward. I am not likely to draw out every detail, but by the time this dissertation is finished, I hope to have placed at least a board over this gap in explanation, such that the way to the other side is clear even if precarious. In what follows, I begin by tackling the weaker combination problems. I will either reveal that they are not really problems, that they depend upon false presuppositions, or that they are subject to various solutions. I will then shift my attention to the subject combination problem.

I argue, perhaps unsurprisingly, that subjects are not irreducible. It turns out that there are a number of ways of making sense of what it means for a subject to reduce (or, from the other side, to combine). I won't side with any particular account of subject combination/reduction, though I will indicate which solutions I am more or less sympathetic to. If so much as a single one of my proposed solutions works, then the subject combination problem is no more. Now, that subjecthood is irreducible is ultimately an intuition, albeit a strong one. This hasn't stopped certain philosophers from claiming to have made such arguments, though, upon closer inspection, those arguments turn out to presuppose the very irreducibility they aim to defend (we will see this later on). Given that there is no defense of the intuition, there is no premise for me to attack. Thus, I will not use formal arguments in making sense of subject-summing. Instead, I will use a thought experiment that aims at tearing down the intuition of irreducibility. I hope that my thought experiment will make intuitive just how subjects might sum. All I need is the possibility.

I end the dissertation by summarizing what I have accomplished and sparing some thoughts on the work that still lies ahead.

Section B: Zombies, Palettes, and Phenomenal Exclusivity

In what follows, I will tackle the weaker combination problems before moving on to the subject combination problem. I'll take these on in increasing order of difficulty, beginning with the zombie problem.

The zombie problem is about conceivability. According to Chalmers, we seem to be capable of conceiving of "a world in which microphysics and microexperience is just as it is in our world, but in which no macroscopic entity is conscious" (2016: 10). In such a world, there are "panpsychist zombies, which are microphysical and microphenomenal duplicates of us without consciousness." Goff, in one of his examples, says that in offering a combinatorial explanation of macroscopic pain, we might "suppose that my severe pain intelligibly arises from the slight pain of trillions of neurons. And yet, for any group of subjects feeling slight pain, it seems possible to conceive of just that number of slightly pained subjects existing in the absence of some further pained subject, whether slightly or severely pained" (2017: 174). This problem is, as I'm sure is clear, the analogue of the zombie problem for classical physicalism. It'll be worth it to take a moment to understand precisely why the zombie problem was such a deep issue for classic physicalist theories.

Mary's Room, the inverted spectrum, and the possibility of zombies all stemmed from the same concern with classical physicalism: it seems perfectly conceivable to find ourselves in a situation where we possess all of the physical information and *still* not know anything about the phenomenal domain. Let's use the identity theory as an example, though any classical physicalist theory will do. The identity theorist, we'll recall, tells us that pain is the firing of

C-fibers. We can then close our eyes, envision the firing of C-fibers, and imagine no pain in the world. Trying again, we tighten our eyelids, and this time we imagine the C-fibers firing, the neurons exchanging neurotransmitters, the moaning and groaning, and yet still we can imagine all that absent the sensation of pain. Well, then we've done it: we have imagined a pain-zombie. We can do this for every neural-phenomenal identity the identity theorist draws until we've got ourselves a complete zombie. But, as the identity theorist is quick to point out, *nothing follows from this*. More precisely, the fact that we can imagine such a scenario does not entail that it is possible. This is in fact the move that many identity theorists make: while zombies are conceivable, they are impossible. This is an interesting back-and-forth between the conceiver of zombies and the identity theorist. As it stands, the claims of conceivability do not obviously add up to an objection, and the identity theorist's response fails miserably to satisfy us, the conceivers. Why is that? It's because the conceivability claim, on its own, is not an objection. It is a question, and one that the identity theorist fails to answer.

The conceivability of zombies is a colorful way to ask: how, exactly, do these physical facts settle the phenomenal ones? The idea is that the world, we hope, is ultimately intelligible. If mental stuff truly does arise, through purely physical means, from physical stuff—if mentality is not brutely emergent—then there *must* be an explanation available. As such, the conceivability of zombies is not an objection, it is a request. Indeed, it is a valid, understandable request. So long as zombies are conceivable, the identity theory has failed to provide a satisfying explanation for the link between mind and body. Furthermore, the reply that no explanation is forthcoming because no explanation is possible is deeply unsatisfying. The explanation is

impossible because the identity theorist makes use of brute emergence. Whatever else may be unacceptable about brute emergence, it necessitates the existence of an explanatory gap, and that alone is deeply problematic.

The same thing is happening in the case of micro panpsychism. That zombies are conceivable is not an objection, it is a question: how do the microphenomenal facts settle the macrophenomenal ones? In other words: how can we make sense of combination? This is a perfectly good question, and one that needs a reply if we are not to face the same fate as the classical physicalist. What's important to note here is that the conceivability of zombies is not raising a unique objection; it is merely a gesture at the combination problem. An adequate solution to the combination problem should result in the inconceivability of zombies. Here is perhaps a more hopeful way of putting things: if there is at least *one* story that can be told that makes it conceivable how microphenomena *add up* to phenomena, then we are in infinitely better shape than the classical physicalists. Given that *one* story is true, zombies, on that story, will be literally inconceivable. I hope to provide more than one story that eliminates the possibility of phenomenal zombies.

Moving on, let us now consider the palette problem. The palette problem is the concern that at the micro level of reality we don't have a broad enough palette of experiences to account for the complexity of experience at the macroscopic level. Chalmers articulates the problem: "There is a vast array of macroqualities, including many different phenomenal colors, shapes, sounds, smells, and tastes. There is presumably only a limited palette of microqualities. [...] How can this limited palette of microqualities combine to yield the vast array of

macroqualities?" (2016: 183). Before continuing, it's worth mentioning that this combination problem is meant to be more problematic for the Russellian monist, who believes that the physical terms that physics employs to refer to the properties of fundamental entities pick out the fundamental phenomenal properties. According to Russellian monism, the intrinsic character of matter is left open by physics, and that intrinsic nature is exactly the kind of stuff that phenomenal properties are built out of; furthermore, the intrinsic properties of matter correspond to those posited by physics. As Goff rightly points out, "the vast range of kinds of consciousness — from colors, to tastes, to the experience of echo-location — must emerge from, as it were, a tiny palette of qualities," and it's hard to see how this could be achieved (2017: 194). Here is an immediate solution to this problem. The view we are considering here is not necessarily Russellian monism (though it is consistent with it). As such, we need not commit ourselves to the palette provided by physics. On our view, microphenomena serve as the causal bases of physical dispositions, but we have said nothing about whether these bases can undergo categorical change—allowing for a more dynamic palette—nor about how many we can expect to find. After all, two distinct phenomenal bases can give rise to the same manifestations (this does not mean that they possess identical dispositions, merely that similar behavior can be observed). Given this, I believe that the palette problem just isn't that worrying. If we reject Russellian monism, there is no reason to restrict ourselves to a limited palette. Furthermore, even if we do restrict ourselves to a limited palette, there's no obvious reason to believe that such a palette cannot give rise to the experiential complexity with which we are familiar. However, the palette problem is popular, and I certainly should say more.

Regardless of whether we are Russellian monists, it certainly seems plausible that the palette of microphenomenal properties will not be very diverse. How, then, could we account for macroexperiential phenomena?

There are at least two problems here that I need to deal with. The first is just the general palette problem as articulated above. However, the second is a species of the first that makes things considerably harder: experiences are modal, that is, we have five sensory modalities.[50] Sights are not sounds, sounds are not tastes, and tastes are not tactile. As such, I should feel extra pressure to provide an explanation for how a limited palette could account for modality. How could microphenomena, which are supposed to be simple, possibly combine to account for modal macroscopic phenomenology? To put things more vividly: there is a tension between the modality of experience and the scarcity of microphenomenal qualities. If we disallow for modality at the fundamental level, then we seem to have too few properties. If we allow for modality, then we seem to have too big a palette. How do we solve this? I believe both of these problems have plausible solutions. I will begin with the broad concern, ignoring modality. I will then address the modal worry. Finally, I will offer yet another possibility that may provide an independent solution to both problems at once.

To reiterate, the broad palette problem is that our palette is too small to account for experiential complexity. I think the most straightforward response here is to deny that this is really a problem. This is one of the options Chalmers believes shows some promise, saying that "small-palette solutions argue that all macroqualities can be generated from just a few

[50] For the sake of simplicity, I will restrict myself to the familiar five modalities, even though there are supposed to be many more. What I have to say should work regardless of the number of modalities there are.

microqualities, if we find the right underlying microqualities with sufficient flexibility and generality" (2016: 207). Generally speaking, that we can generate a great number of arrangements, combinations, possibilities, etc. from a small number of available resources isn't normally surprising. Consider an example. With the colors cyan, yellow, and magenta, we can generate the entire color spectrum. That's roughly 16 million shades. These are the colors used by standard printers. At the end of the day, this is just an analogy, and what's worse is that it directly appeals to color. But consider physics. Indeed, consider the objection as it is posed to the Russellian monist. Physics, it is said, posits *too few* microphysical properties, thereby resulting in there being too few microphenomenal properties to account for the complexity of experience. Why would we think this? Physics posits the number of microphysical properties it does precisely because they suffice to explain the *immense complexity* of the universe. From a handful of microphysical properties, physics aims to explain everything from the interactions between quanta to the Hawking radiation of black holes, the interactions between asteroids and planets, and the movements of neurotransmitters in our brains. Furthermore, physics is likely to posit more fundamental entities as it attempts to explain the existence of dark matter and the like, and this is to say nothing of antimatter. So, my response to the broad palette problem is two-fold. First, it is far from obvious that the palette is too small. From a small palette, physics has explained great complexity. So from a small palette, the great complexity of experience may also be subject to explanation. Second, it is extremely unlikely that physics is done positing fundamental properties, such that if we are restricted to working with the properties that physics delivers, it's likely too soon to know what size of palette we're dealing

with. It isn't clear to me why the phenomenal domain would have a combinatorial restriction that doesn't seem to appear in any other domain.

But what of modality? To make the modal objection salient, Goff asks: given that "the taste of mint on the one hand and the experience of red on the other have nothing whatsoever in common," how could they be built up of the same kind of microphenomenal stuff (2017: 195)? The fact that experience is modal certainly suggests that we will need more than just one or two more properties, no? I don't know that this is true, but we do have a response available: experience is not modal. Certainly, *if* experience is not modal, then the reply to the broad palette problem is all we need. However, we now need to make clear what it would mean for experience not to be modal, given that it certainly seems to be!

Here is what it would mean for macroexperiential phenomena not to be modal. Goff cites an early draft of Coleman's "Neuro-Cosmology" paper in which Coleman suggests that our experiences lie on a single spectrum. In that earlier draft, he muses about an experience that most of us have probably had at one time or another in which he wakes up in the middle of the night to what was either a loud sound or a jolt (the kind we feel if the bed is suddenly moved). He is certain that it wasn't both, but he cannot tell which one it was. Goff says that perhaps "we think there is an unbridgeable gap between colors and tastes simply because we lack the experiences that would bridge that gap. Perhaps there is a certain range of possible experiences (not had by humans) that lie in between auditory experiences and color experiences, such that if we instantiated those 'in between' experiences we would be able to move in imagination from colors to sounds as seamlessly as we move between shades of blue" (2017: 200). I find this

highly plausible. Indeed, take two of our sensory modalities: taste and smell. They are remarkably similar! Indeed, they are so close together that when a smell is strong enough, I'll think or say something like "I can almost taste it." It's as though taste is actually just a very strong smell, one that's had on the tongue rather than in the nose. What Coleman was getting at in the passage Goff engages with is that perhaps sound and touch are also not as far apart as we think. Perhaps there are possible experiences that we do not have—that, in virtue of being human, we are incapable of having—that could reveal that the two lie on the same spectrum. If this is right, then it turns out that the experiences that we take to be modal are actually just sub-ranges within a broader experiential spectrum. Without having to account for genuine sensory modality, we could make use of a more restricted palette.

I wish to provide a simple thought experiment that I hope will make the possibility of a continuum of experience more palatable. Imagine that there exists a creature whose visual experience (felt through the eyes of the creature in the same way our visual experience is) is entirely in grayscale. Furthermore, and this is where we must stretch our imaginative muscles, the creature has a sense of touch, but it is not tactile like ours. Instead, this creature, upon touching surfaces, experiences bluescale responses that vary in saturation and brightness depending on the surface touched. This bluescale experience is experienced *at the fingertips* (or whichever part of its body it uses at that moment to engage with its surroundings), not within its field of view. Thus, same as we experience a tactile response upon touching a surface with our fingertips *at* our fingertips, this creature experiences a bluescale response at the location of contact. Such a creature would surely, like us, believe that its grayscale sight and bluescale

touch are different, incommensurable modalities. But it would be wrong. I take it that such a possibility is conceivable. I find it reinforced by the similarities between some of our modalities. If there are no experiential modalities, only the illusion of such, then that they could be built up of a small palette becomes unproblematic. But we needn't even go that far. Perhaps there are modalities, but this possibility at bare minimum suggests that there may be fewer than expected. If it turns out that there are only two, I believe that a small palette could suffice. Thus, I have two replies, one strong and one weak. The strong one is that there are no sensory modalities. I am partial to this. The weak is that there are likely fewer sensory modalities than we believe. Either way, all we need to respond to the modal concern is that this be possible. If it is possible, then there is no palette combination problem, for there is a conceivable way forward. It's certainly nothing that amounts to an explanatory gap.

There is one final solution that the reader may have already noticed we have available. I have been engaging with the palette problem assuming that our microphenomenal palette exhibits Weak Character. It is precisely by assuming Weak Character that we may have the worry that there just isn't enough experiential paint with which to draw the universe. Given that I want both phenomenal characterizations to be viable, it was worth the expended ink to address the palette problem in this way. If, however, our palette exhibits Strong Character, then there is certainly no palette problem. A palette with Strong Character has every imaginable resource at its disposal.

Let's now consider the final problem of this section: phenomenal states are exclusive. Coleman believes that he can prove decisively that combination is impossible, and, worse,

incoherent. Given the ambition of his goal, it will be unsurprising if he falls a little short. The problem, says Coleman, is that phenomenal states (including microphenomenal states) are as exclusive in their contents as they are inclusive. This means that, for example, the sum total of my experience is characterized not only by what is in it, but also by what is not. Let's consider the problem the way he formulates it. He offers us a straightforward thought experiment, and I feel that it is best to preserve it, so I shall quote it in its entirety:

> Consider the original duo's point of view. One—Blue's—is pervaded by a unitary blueness, the other—Red's—by redness, and that is all they experience respectively. To say these points of view were present as components in the experiential perspective of the uber-subject ("Ub") would therefore be to say that Ub experienced a unitary phenomenal blueness and a unitary phenomenal redness, i.e. had synchronous experiences as of each of these qualities alone, to the exclusion of all others. For it is these qualities each on their own that characterise, respectively, the perspective of the original duo. Experience excludes, as well as includes. Yet nowhere does Ub have any such experiences: he precisely combines his predecessors' qualitative experiential contents. Ub doesn't experience red-to-the-exclusion-of-(blue-and)-all-else, nor blue-to-the-exclusion-of-(red-and)-all-else, let alone—impossibly—both together. Thus, the original points of view are not ingredients in Ub's subjectivity. (Coleman, 2013: 33)

To break things down, the problem seems to be that Blue's experience is blue and only blue, whereas Red's experience is red and only red. Upon combining, the subject they compose, Ub, has only so many options for what it experiences, and the options all seem bad. Ub's experience might be solely red, in which case Blue contributes nothing and we have no combination. Ub's experience might be solely blue, in which case Red contributes nothing and we have no combination. Ub's experience might be half-red and half-blue, in which case neither Red nor Blue comprise Ub's experience, as their experiences are exclusive. Finally, Ub's experience

might be both Red's and Blue's, but this would result in a contradiction due to the exclusivity of the experiences of Red and Blue while combined.

Ub, it seems, has no hope of existing: its parts are simply incapable of combination. But why is this? Combination is not some special process that's unique to minds. The term "combination" appears to be a neutral way of picking out any process of composition. We know that things other than minds can combine, and everywhere else, this process of combination doesn't seem all that mysterious. Let's consider a more straightforward instance of combination. The following seems to be clearly true: when I combine four wooden legs, a wooden seat, a wooden backrest, and a pair of wooden armrests together in the correct arrangement, the result is a wooden chair. This isn't to say that there's nothing mysterious here—composition is notoriously difficult to make sense of. However, no one would say that what just occurred is impossible. Furthermore, there are numerous accounts of what occurred in the case of the chair. Let's briefly consider two.

The first account has it so that upon combining the parts of the chair, we get a new thing: the chair. This new thing is the product of its parts being arranged as they are. Thus, once we finish counting all of the things present where the chair is located, we will count nine entities: four legs, one seat, one backrest, two armrests, and a chair. Could the chair have totally different properties from its parts? Of course not. The chair is entirely dependent upon its parts, it is made from them. Its color will be determined by the color of the parts, as will its degree of comfort, sturdiness, etc. The chair may be a different *thing*, but it is intimately tied to

its parts. Arranging these parts together would not have given rise to a steel rocket—the result must be a wooden chair.

The second account disallows the creation of any new entity (e.g., a chair) upon the combination of its parts. Thus, when we've arranged the various parts, what we have in the end is precisely just those parts arranged as they are. But we certainly speak of things like chairs as objects, and they are objects of which certain statements seem to hold true. To make sense of this, the account we're considering claims that truths that appear to be about the chair are made true by the properties of the parts and their arrangement. For instance, that the chair is comfortable is true because of the seat and backrest.

I greatly prefer the second account above, and so I will largely default to it, though I will make explicit when I deviate from it. All the same, either account will suit my purposes. The problem with Coleman's objection isn't which type of composition he is partial to, but rather that he eliminates any possibility of phenomenal composition through a dubious assumption. Let's turn to that now.

Let's take a much closer look at the above thought experiment concerning Ub. The way things are presented, the stage is set such that when the curtain is raised, we are presented with the following: here is Red, experiencing only-red; one micron away is Blue, experiencing only-blue; finally, Red and Blue, we are told, have already done whatever is necessary for successful combination, so we have Ub. And so we are doomed to fail. Regardless of which account of composition we opt for, we will not get what we want. If Ub is an entity in its own right, it has the impossible choice articulated above: only-red without combination, only-blue

without combination, or the contradiction of accepting both. If, alternatively, all there is is Red and Blue arranged as they are, there are no facts about them given how they are presented that would give any reason to believe that combination has occurred. We might as well place Red and Blue several lightyears apart. But it seems that the thing we are being asked to imagine begs the question.

Coleman's objective is to prove that combination is impossible. Well, what if we change the scene a bit? Couldn't we grab Red and Blue prior to combination, and then, upon combining them, say that their experiential contents changed? For instance, one way this might occur is by, once properly combined, having Red's experience change to half-red/half-blue and Blue's experience undergo the same change. As an alternative, Red's and Blue's experiences might become purple.[51] If this happened, Ub would have no impossible choice to make. Ub's experience could be half-red/half-blue or purple, and Red and Blue would be accounted for. Coleman believes this cannot happen, and that is because of the exclusivity of experience: the experiences of a subject must exclude all other experiences. Must experience be exclusive in this way? Well, it isn't obvious exactly what Coleman means by 'exclusive' in this context, but there's only option that can get his point across. If, for instance, 'exclusive' just means that there are things that the subject is not currently experiencing, then I am in agreement, but this does not grant Coleman the result he wants. That experience is like this shows only that at any given point, there are facts about what is excluded from a subject's experience. So, perhaps prior to combination, it made perfect sense to say that Red's experience excluded Blue's.

[51] We will explore these possibilities in greater detail in the next section.

However, this does not necessitate that the experiences of Red cannot change, nor does it give us reason to believe that combination of Red and Blue is impossible. For instance, perhaps when Red and Blue are within one micron of each other, their experiences combine and become purple.

Coleman can deny this, but to do so is to assume that combination is impossible, not to prove it. His thought experiment depends upon this extreme exclusivity, and it is not something he argues for—he merely takes it for granted. Consider the fact that Coleman built in to his thought experiment that Red and Blue are already composing Ub, yet their singular experiences are mutually and necessarily exclusive. Coleman is presupposing that subjects necessarily exclude the experiences of one another—he is presupposing that they cannot combine. In other words, no matter how tightly you squeeze Red and Blue together, you will not get their experiences to blend. They *must* exclude each other. And so we arrive at the heart of Coleman's Ub case. The intuition that's driving Coleman's thought experiment is precisely that subjects are units that can neither decompose nor combine. Without this intuition, nothing seems to stop us from saying that the experience of Blue, upon joining Red, now includes redness. Perhaps when combination occurs, the experiences of the micro-subjects change. It might turn out that Coleman's thought experiment is impossible: if you place Blue and Red together in the right way, they cannot have blue-only and red-only experiences, as this would just be to deny combination without argument.

If Coleman is right and Red, Blue, and Ub all turn out to be subjects and subjects are neither decomposable nor combinable, then the thought experiment seems unnecessary—it

turns out, as a matter of definition, that the combination problem is not resolvable. Clearly, there's something that's gone wrong here. However, the issue we are now dealing with is the subject combination problem. Thus, in order to make sense of the experiences of Ub, Red, and Blue, we'll need to solve the subject combination problem. In what follows, I aim to do just that. I will begin by putting all of the intuitions about subjects on the table so that we can inspect precisely why we might believe that subjects are irreducible. Both Goff and Coleman claim to offer arguments, though their arguments ultimately turn out to be deeply held intuitions (nothing wrong with that!). However, I will argue that there are serious mistakes made in the way they characterize subjecthood. I'll provide a better characterization, and I will then show what it means for subjects to combine.

Section C: The Subject Combination Problem

We must now turn our attention to the subject combination problem. The problem, of course, is that it is extremely difficult to conceive of how subjects may combine. This sentiment is sometimes expressed weakly as a gesture at the opacity of a solution, and sometimes strongly as a conviction that a solution is impossible. The intuition that subjects can neither reduce nor combine is so powerful that no formal arguments have truly been offered—it is mostly taken for granted. Though, as we will see, attempts have been made. I wish to proceed as follows. First, I want to provide a simple definition of 'subject'. This simple definition will be subject, pardon the pun, to revision later on. Or, at least, to greater precision. After providing the definition, I wish to wade through the murky waters of the intuitions against subject-summing

and subject reduction. I will, for the most part, refrain from objecting and instead build these up as much as possible. I will, on occasion, comment on peculiarities or relatively minor mistakes. Afterward, I will reveal a multitude of deep problems with these anti-reduction intuitions, all of them stemming from unjustified modifications to our definition of 'subject'. We will then turn our attention to what I believe the greatest mistake is in this debate: misunderstanding what it would mean for subjects to combine. Finally, I will share some thoughts, with Goff's blessing, on what we need to do to show that combination *is* possible. It is at this point that we will reach the positive part of this section. I will provide a thought experiment that I believe challenges the anti-reduction intuition. The remainder of this section will be dedicated to making sense of what we should take away from that thought experiment. Indeed, if all goes well, it will tell us what we should think about Coleman's Ub, zombies, and combination generally.

I propose the following definition of 'subject': to be a subject is to experience. This definition is largely in line with what Goff and Coleman conceive, and given that they are both adamantly opposed to the combination of subjects, I take it that it will be a fair one. Goff defines 'subject', as mentioned previously, as a determinable of conscious states. He says, to recite his example, "to be pained is to be a subject in some specific way; to have an experience of orange is to be a subject in some other way" (Goff, 2017: 178). Coleman offers a slightly different definition, saying only that "a subject is just that sort of entity for whom anything can be like anything at all" (Coleman, 2013: 25). These definitions are importantly different. Goff believes that *all* instances of experience are experienced by subjects. In his view, it is impossible

for there to exist an unexperienced experience. If there is a pain, there must be something that feels that pain. Coleman disagrees. Coleman is what he calls a 'panqualityist' rather than a panpsychist. He believes that qualities can most certainly exist without being experienced. We will ignore these differences, as what matters to us isn't whether experiences can exist without being experienced, but rather what subjects are, and on both views subjects amount to the same thing. That is, they are both in agreement that to be a subject is to have experiences. Let us now turn to the multitude of reasons to believe that subjects just aren't the kind of thing that could ever combine or reduce.

First and foremost, that subjects cannot reduce is taken as beyond question. For instance, Goff says plainly when discussing the impossibility of subject-summing, "the starting point is just a deep intuition that *subjects aren't combinable*" (Goff, 2017: 171). He asks whether there is anything that we know about the human consciousness that makes it obvious that it isn't composed of smaller subjects, and when asked in this manner, he believes "we do meet insuperable challenges," in fact "we have very good reason to think that subjects are irreducible" (Goff, 2017: 209). It is worth mentioning here that no such reason is ever provided. Instead, as will become apparent, the "very good reason" is that it's hard, or perhaps he would say 'impossible', to conceive of.

At times, it appears to almost be a matter of definition that subjects cannot reduce. Goff says that "it doesn't seem that we can specify what it is for there to be a conscious subject in more fundamental *terms*" (Goff, 2017: 20, *emphasis mine*). Here, the issue seems to be that something about the meaning of 'subject' bars a decompositional analysis. Indeed, were we

providing an a priori analysis of subjecthood, "then it must be in some implicit sense what we *mean* when we judge that there is a conscious subject" that it is composed of smaller subjects (Goff, 2017: 212). But, he goes on, "it is simply not plausible that my judgment that there is a conscious subject [...] consists in the judgment that there are a large number of micro-subjects, none of which is identical with [the macro-subject], standing in some relation." He doesn't consider that perhaps 'micro-subject' and 'macro-subject' could have different definitions. Perhaps macro-subjects turn out to be composed of micro-subjects, and that is something we can discover about the world. We could pick out macro-subjects by ostension, without a thought spared to what they actually are.

Sometimes the intuition that subjects cannot reduce or be composed relies on the perceived unity of our daily conscious experience. This should be a familiar worry, as it is undeniable that we conceive of ourselves as units. Coleman's argument against composition ultimately bottoms out on this intuition. He asks that we consider a particular experience, one in which we are "cold, tired, and smelling roast beef" (Coleman, 2013: 23). In such a scenario, our "phenomenological point of view appears suffused by all three [experiences] together, as opposed to experiencing them only discretely, in series." Since we cannot see any individuated parts in our subjective experience—since everything 'blends' together—it is clear that we are a unit. Coleman takes this intuition as part of what subjecthood is. It comes as no surprise, then, that subjects cannot combine or be decomposed; they are, by definition, simple units. This is why Ub runs into trouble. Of course, this shows once again that he presupposes that which he aims to prove, but we'll come back to this later.

The last intuition we'll consider for now (though there are many others) is that surely our experience is different from that of our composing parts. Goff has an example he likes to keep coming back to throughout his publications concerning instances of pain. He asks that we "suppose that each of the billion ultimates that compose my brain is a subject of [...slight pain]. It is unintelligible why the arrangement of these ultimates [...] should give rise to some *new* subject of experience, over and above the billion slightly pained subjects of experience we already have" (Goff, 2006: 54). There is a lot going on here, some of which I'll address shortly. For instance, the existence of a truly novel subject appears to directly contradict the micro panpsychist's claim that the macro-subject is a combination of the micro-subjects. However we decide to make sense of combination, it cannot be the case that a truly new entity appears that is wholly independent of its parts, as that would just be brute emergence once again. We'll get back to that in a bit. What we should focus on here is that the macro-subject, *us*, feels something different from the micro-subjects. So, we are something else from the micro-subjects. Notice that this falls well short of an argument, it is merely a reassertion of the intuition that subjects cannot decompose or combine. If they could, the example would be presented quite differently.

These intuitions all share the same thing in common: we (referring to those like Goff and Coleman) cannot conceive of what it would mean to combine. As such, every example has a lack of combination built in. Every example mirrors the original one provided by William James, which possesses a mistake, and which we will discuss. For now, what's important is that these intuitions be out in the open for what they are: *intuitions*. I would like to note that the

problem isn't that they are intuitions. Rather, it is that they are intuitions presented as arguments. This matters because it affects how best to respond to them. Let us now turn our attention to some of the mistakes that are being made. Specifically, these mistakes are mistakes of building in too much to our notion of 'subject', the notion we considered at the start of the section.

The first mistake that seems quite prevalent is the conflation of 'subject' and 'self'. A self is meant to have a psychology and have a sense of cohesion; we take our*selves* as paradigmatic instances of selves. But a subject, as I've already said, is far less robust. That this mistake is made is quite understandable. *We*, presumably, are selves, but we are also subjects. Nonetheless these two concepts come apart. Consider what Coleman says about roast beef, tiredness, and being cold. That we feel that these things are all held within the same entity follows from the fact that we believe that we are selves. Selves are supposed to be units. Indeed, that is built in to the definition of 'self' (rather, into any *attempt* to define 'self'; as Nagel makes very clear in his book, *The View from Nowhere*, selves are exceedingly difficult to pin down). Coleman says when detailing his example that the experience "is *unified* in the following sense: though the subject can attend now to the sensation of cold, now to her tiredness, and now to the smell of roast beef, still, phenomenologically-speaking, these three sensations are given to her all in one go" (Coleman, 2013: 23). That all of the sensations happen at once to the same self need not say anything about the unity of subjecthood. Now, one might rightfully object that these are all being perceived at once by one thing. Perhaps. For now, I simply wish to point out that 'self' and 'subject' are different. Electrons may very well be subjects, but they are not

selves. That *we* feel unified may be nothing more than a symptom of being a self or having a particular psychology. It need not follow from the fact that we are also subjects. In fact, many Buddhists deny that we possess any kind of unity. Perhaps simpler entities that are not selves can have disunified subjecthood.

Returning now to the objection just raised above, isn't it obvious that when we look within ourselves we find just one subject? Upon introspection, try as we might, "a vast array of microexperiences is not revealed to us…" (Chalmers, 2016: 190). Surely, the intuition goes, were we made up of micro-subjects, we would *see* the micro-subjects. Instead, we see only the macro-subject. My response to this is simply that this is not obviously the case. Let us ask: assuming we *were* made up of smaller subjective entities, what would it be like to see them? Should we expect to find the edges of the pixels of our visual experience? Should we feel all one-hundred billion experiences *as* one-hundred billion experiences? This is baffling. When I look within myself, what I see does not strike me as obvious unity. In fact, there's a sense in which experiences are so disunified that I am psychologically convinced that I have *five* modalities of sensation! Perhaps the relevant sense of unity comes in when considering that those five modalities are all *my* modalities—*I* experience them all. However, maybe this a contingent matter. It could be that infants don't feel this unity or sense of *I* at all, and it is something that gets built up over time as we learn how these senses relate to the external world or as we develop a sense of self.[52] Furthermore, I argue with myself, feel myself pulled in many

[52] I don't believe that selves are the result of subjecthood. Perhaps selves admit of functional analysis or are some purely psychological entity. Maybe a self is nothing more than some set of beliefs. I don't quite know what to make of selves, but I am not committed to the view that subjecthood somehow adds up to selfhood. It may turn out that we can have macro-subjects that are not selves.

directions, experience conflicts, drive on autopilot whilst thinking of dinner, and so on. My response here is not that these experiences prove we are made up of other subjects, rather that the objection that we feel unified holds precious little water. The fact is, we have no idea what unity or disunity should feel like. As such, it just isn't true that we should expect to see the parts making us up. I can think of no reason to believe this. But there's more to be said. It simply isn't true that we experience anything approaching a subject when we introspect. Consider my thoughts above about the self. It is *from* the notion of 'self' that we believe in unity. But as Nagel is quick to point out, we can find no such self upon introspection: "The apparent impossibility of identifying or essentially connecting the self with anything comes from the Cartesian conviction that its nature is fully revealed to introspection, and that our immediate subjective conception of the thing in our own case contains everything essential to it, if only we could extract it. But it turns out that we can extract nothing, not even a Cartesian soul" (Nagel, 1989: 34-5). We look within and find sensations, thoughts, vague feelings of emotion, etc. Nowhere will we find that which holds them all. It was this very fact that led Hume to claim that we are nothing more than a bundle of experiences. The same holds for subjecthood. However many subjects we turn out to be, I submit our actual experience should not sway us in any direction. Again, too much is being built in to our simple definition of 'subject'.

While I believe there are more errors in this debate—some of which we shall consider in what follows—there is one final mistake that I think drives the central intuition against combination. The best way I can articulate it is like so: subjects have *skins* or *shells*. Take as an

example what Coleman says about subjects being experiential entities. He tells us that being a subject can be seen as being "a discrete 'sphere' of conscious-experiential goings-on [...] with regard to which other subjects are distinct in respect of the phenomenal qualities they [the original subject] experience, and they have no direct (i.e. experiential) access to the qualitative field enjoyed by the first subject" (Coleman, 2013: 30). The first thing that should strike us about this characterization of subjecthood is how it guarantees that subjects are irreducible. If it is built in to the definition of subjecthood that subjects are "discrete spheres" that "have no direct access" to the "qualitative fields" of others, then the remainder of what Coleman says against the combination of subjects becomes superfluous. It turns out that it is *in the very nature of* subjecthood that it is irreducible. If only I could avail myself of a similar style of argument to prove my point!

I would like, now, to provide an alternative conception of subjects. I do not wish to *keep* this characterization, only to offer it as an alternative *for now* in order to reply to the supposed discreteness of subjects. First, I wish to issue the following reminder: Coleman is a panqualityist. This means he believes qualities (of the qualia kind) are ubiquitous throughout nature. It is subjects that are not.[53] Recall further that we have a fairly simple definition of 'subject' on board. Physics tells us that electrons, under one characterization, are disturbances in the electron field. Now, electrons are sometimes characterized as being non-extended points. Nonetheless, regardless of whether they are disturbances in the electron field or points that generate fields, those fields *do* take up space, and this suffices for extension.[54] Furthermore,

[53] Notice how sharply this contrasts with what Goff would be allowed to say.
[54] See Strawson (2006): 16.

those fields *interact* with the fields of other particles. Indeed, they overlap. Here is one way of thinking about microphenomenal causal bases. They take up the amount of space that the electron's field does. If so, then these fields, upon interacting, would result in the microphenomena of both of these entities literally blending together. If this sounds a bit far-fetched, it is now that it is worth mentioning that this very possibility has been defended by Coleman himself. He states that "we might envisage, for example, the fundamental physical world as a continuous but variegated quality field, or several intersecting quality fields, like enormous sheets of different colors" (Coleman, 2015: 86). Seager, too, suggests a way we might think about the combination of subjects in this manner by appeal to the superpositioned states of particles in the double-slit experiment (Seager, 1995: 284). Coleman, being a panqualityist, believes this kind of blending is perfectly acceptable when it comes to qualia, just not with subjects. But this rests on his presupposition that subjects are these discrete spheres. If subjects turn out to be more in line with how physics draws our particles—if microphenomenal subjects extend along their fields—then combination, blending, and so on is back on the table. I don't believe this is the only possible way of allowing subjects to merge; the objective is simply to show that the intuition that subjects are "shelled" or "skinned" is not one we need to accept. If subjects are not discrete in the way Coleman makes them out to be, then the possibility of combination is still within reach.

As a final note that we need not take too seriously, skins and shells are ontologically dubious. If we are positing microphenomenal properties at the fundamental level of reality as the causal bases of dispositions, there is nothing to separate them from the outside world.

There is no shield, no shell, no skin, and no other *thing* that necessitates that they cannot overlap. This appeal to the barriers that separate the internal life of the subject from the external world is likely motivated by our belief that our mental lives are inherently and necessarily private. But the privacy of experience and the existence of mind are two separate issues, only the latter of which gives rise to the explanatory gap. Perhaps there is some other way of preserving the walls that separate our macrophenomenal mental lives from those of others that does not require a similar barrier at the microphenomenal level. But it is the existence of macrophenomena which I must make sense of here, not privacy, and if I must put the sanctity of the isolated status of mind at risk in order to close the explanatory gap, then my aims are sacrilegious. Whether we can make sense of combination is one thing, but that it is a consequence of the concept 'subject' that we can't is false.

And so we return to the definition of 'subject'. I maintain that a subject is nothing more than that which experiences. There is more to be said, but what I hope to have shown is that it does not follow from the meaning of 'subjecthood', nor of metaphysical necessity, that subjects be the types of things that cannot reduce or compose.

There is one more problem we must consider before we move on to my positive account of combination, and that is the original iteration of the subject combination problem, famously articulated by William James. Here it is in its entirety.

> Take a hundred of them [feelings], shuffle them and pack them as close together as you can (whatever that may mean); still each remains the same feeling it always was, shut in its own skin, windowless, ignorant of what the other feelings are and mean. There would be a hundred-and-first-feeling there, if, when a group or series of such feelings were set up, a consciousness *belonging to the group as such* should emerge. And this 101st feeling would be a totally new fact; the 100 feelings might, by a curious physical

> law, be a signal for its *creation*, when they came together; but they would have no substantial identity with it, nor it with them, and one could never deduce the one from the others, nor (in any intelligible sense) say they *evolved* it. (James, 1890/1981: 160)

Goff considers this a "spelling out" of the subject combination problem, though it's just a restatement of it (Goff, 2017: 172). Before we get down to the main problem here, let's take a moment to appreciate the presence of some of the intuitions we have already addressed. We see in this quotation a reference to the 'skin' of the composing parts. Furthermore, due to this skin, the components cannot be 'aware' of one another. And the creation of this 101st subject that is entirely novel echoes the deeply held intuition that the 100 subjects cannot *combine* to form the macro-subject. It is this last bit that I wish to address.

As I have mentioned before, any account of combination should reject the creation of an entirely new and wholly separate entity upon combining for the same reason we would reject this in the case of the humble chair. There is no additional, independent entity upon combination any more than there is an additional entity once I've put the legs, seat, armrests, and backrest of the chair together. One reason given for believing that there *must* be a *new* thing has to do with, as I've mentioned before, the qualitative difference between the macro and micro experiences. Goff says that the panpsychist *must* assume the experiences to be different, that is, "the experiential being of a higher-level subject of experience is significantly qualitatively different from the experiential being of the lower-level subjects of experience of which it is constituted" (Goff, 2006: 57). But this is wrong. The only thing the panpsychist must grant is that the experiences of macro entities must differ from those of micro entities. It needn't be the case that the experiences of *this* macro entity differ from the experiences of *this*

entity's constituents. Again, what this means is something we'll explore shortly. What matters for now is that the driving intuition is mistaken. The panpsychist can hold that micro and macro experiences differ without holding that this is true *of one entity and its parts*.

To provide an example for the above which I ultimately believe can't work for unrelated reasons, one might think that micro-subjects *fuse*.[55] If micro-subjects fuse, then there is no 101st subject to speak of that is independent of the other 100. There is just the one subject and the one experience, since the previous 100 merged together into a single subject, losing their prior individuality. In this instance, there are no micro-subjects to speak of. This would be a clear instance of composition where it does not make sense to claim that the macro experience is different from the micro experiences of the entity's composing parts.

Chalmers offers a diagnosis of this seemingly unshakeable intuition, and he suggests a way out. He believes we might be presupposing that subjects are primitive entities. If so, then "they could not be constituted by more basic entities, and combinatorial views would be ruled out" (Chalmers, 2016: 198). In order to make progress, we could, he says, deny that subjects are metaphysically primitive. I agree that it is this very assumption that is being taken as an argument against combination, though I don't think the solution needs to be a denial of the metaphysical primacy of subjecthood. Perhaps we can deny that macro-subjects are primitive; indeed, I believe we should. However, micro-subjects can be taken as primitive without generating any combination problems.

---

[55] This is defended by Mørch (2014) and Seager (2016).

Let's assume that my preferred account of composition is correct: combination does not give rise to any new entities—after combination, you still have only the parts and their arrangement. If the relation between the macro-subject and the micro-subject works in this way, then James is wrong to speak of this 101st subject. There is no true macro-subject in the same way that there is no true chair. Instead, truths about the macro-subject are made true by the experiences of the 100 subjects. Consider Goff's pains. Perhaps when the micro-subjects appropriately combine, their intrinsic characters are altered such as to make sense of truths about intense pain. Coleman offers one way this might work that we'll see in the next paragraph, and we'll delve into greater depth as to what this might look like further on as well. Alternatively, we could take the other account of composition. James, then, would be right to speak of the 101st subject, but there would be plenty of reason to believe that the 101st subject's experience is that of its composing parts. Just as wooden chair parts give rise to wooden chairs and not steel rockets, pain parts should give rise to subjects of pain. To conceive otherwise is to conceive that composition—combination—has failed.

Coleman offers us some ways of thinking about composition, though, of course, he believes this story could not be told of subjects. According to Coleman, "combination [...] is the formation of a whole from components where the components continue to exist in the whole, but are intrinsically altered by combining with one another" (Coleman, 2013: 30). He believes that something goes wrong with subjects, as "we want the ingredients to survive in the whole" when combining, and just "as atoms are not obliterated, only deformed, in building a molecule, subjects would have to persist *as such* within their higher-level product" (Coleman,

2013: 32). The double-standard is interesting. Atoms, for whatever reason, are allowed to persist in a deformed manner, but subjects must continue "*as such*." There cannot be a whole, claims Coleman, because the existence of the macro-subject would require that "at least one subject has gone out of existence, which is not combination but a fight to the death" (Coleman, 2013: 32). This is peculiar. It is, I should mention, unclear what Coleman has in mind when he speaks of the deformation of atoms; whatever it means, I don't see why we wouldn't be able to say something similar about subjects: that is, it isn't clear why we can't say of the macro-subject the same thing we can say of the molecule or why we can't say of the micro-subjects what we can say about the atoms. When looking at H_2O, there is a molecule, but its two hydrogen atoms and the one oxygen atom are all still there. On one account of composition, there aren't four things, just the three. This does not mean that there exists H_2O, hydrogen, hydrogen, and oxygen while at the same time the statement "there are three things" holds true. That would be a contradiction. Rather, what it means for it to be true that there exists an H_2O molecule is just that there are three atoms appropriately arranged. There were three things prior to combination, and there are three things after combination, but, and this is vital, those three things have been *modified* after combination has occurred such as to warrant the use of the term 'H_2O' when referring to them together. Why not say the same about the subject? The macro-subject is the analogue of H_2O, its micro-subjects are the hydrogen and oxygen, and the micro-subjects have combined and appropriately "deformed" so as to constitute the macro-subject in the same way the hydrogen and oxygen do. Regardless of which

account of composition we opt for, micro-subjects will have to undergo some experiential change when combining so as to account for the experiences of the would-be macro-subject.

Here's one reason those who agree with Goff and Coleman might give for denying that the same thing can be said about the subject. In the case of H_2O, it is clear how we get our ontological free lunch. The H_2O molecule is nothing over and above the combination of hydrogen and oxygen. This is much harder to make sense of in the case of subjecthood. In talking about composition and over-and-aboveness, Goff speaks of the peculiar nature of some facts being nothing over and above others. I admit that I don't like thinking about this relation in terms of facts, which I take to be linguistic entities, but I shall follow him here. Goff asks "how can fact X involve different objects and properties to fact Y, and yet, from the perspective of serious metaphysics, add nothing beyond the objects and properties already involved in Y? Philosophers trading in 'nothing over and above' talk owe us an account of how they get their free lunch" (Goff, 2015: 382). Right before saying this, Goff picks out the example of a crowd being neither identical with nor distinct from its members. I find this puzzling. Perhaps crowd-ness is not identical with the members of some particular crowd, but *this* crowd is certainly just its members. My response to Goff is this: I did not get a free lunch. I paid for the bread, ham, swiss, lettuce, tomatoes, and condiments, and I paid for them to be arranged as they are. To then say that I paid for these things but my lunch itself was free is bewildering. Nothing is free. That includes ontology.

By now I hope to have made the following clear. It is not the case that we need to believe that the combination of subjects generates new independent ones, that subjects have

skins, that we are clearly not a multitude of subjects (or the opposite; neither is clear), that subjectivity and selfhood are the same thing, or that subjects are obviously, intuitively, definitely, or otherwise indivisible. The other thing that should be clear is that the claim that subjects cannot combine is an intuition. It is not the conclusion of an elaborate argument, nor does it clearly follow from unobjectionable premises. I wish to further say that the intuition that drives the claim that subjects do not combine is a perfectly understandable one to have. Nonetheless, we must now provide a story of combination. Exactly how much should we do to be successful? Goff believes it would be unfair to ask that I provide necessary and sufficient conditions for combination. I appreciate the kindness, and I will certainly fall short of such an ambitious objective. Instead, Goff believes that "it is not unreasonable to demand some kind a [sic] gesture toward what is required" for combination, adding that "in the case of subjecthood, we're not even able to gesture at its supposed deflationary analysis" (Goff, 2017: 215). It is this impossible gesture that I aim to make.

We must now wade out of the murky waters of intuition and wade into the turbulent waters of speculation. In what follows, I offer the reader a thought experiment. On the basis of that thought experiment I hang most of my hopes. My only request is that the reader give me a fair shot.

Suppose that I approach you with a proposition. Due to miraculous advancements in neuroscience and cybernetics, we have managed to devise the first brain-augmentation chips in history. Their promise is simple. Through a series of relatively non-invasive brain surgeries, we will install these small chips directly into your brain matter. These chips greatly improve your

mental abilities. Indeed, you are promised a 100% improvement over your previous cognitive capacities. Which capacities get improved depends on which chips we install where. You agree, sign a multi-thousand page contract without much thought, and so the process begins. We first install a visual processing chip to your occipital lobe. After the surgery, you notice an immediate improvement in your visual processing. Your depth perception is heightened, colors are crisper, your perception of movement is more precise. You see the world as never before. Indeed, it feels as though you are seeing for the first time. After a month or so, you come back for the next enhancement. We place the next chips—an auditory and spatial processing enhancer—directly to your motor cortex and temporal lobes. Again, you immediately notice a difference. Sounds are more robust and the like, but you also notice subtleties you had never caught before. Furthermore, your spatial awareness and kinesthetic prowess is greatly improved. Your balance is enviable, and you notice that you learn new mechanical skills with great ease. And so we continue until, finally, we aim to improve your reasoning. We install the final chip directly on your prefrontal cortex. Upon waking up, you notice that your thinking is much improved. You can reason your way through complex mathematics like never before. You begin to form novel thoughts, ideas that were once incomplete you now find the solutions to. You are deeply satisfied with the neurological enhancements we have given you.

After a couple of years, you come in for your final checkup to ensure that things are going well. The project has been a success. We remind you at this point that you had signed a consent form that you, in your excitement, only managed to skim through. In that consent form, we had written that we reserved the right to lie to you. However, the time has come to

reveal the truth. We take you into our laboratory, where, beneath a blanket lies a mysterious box-shaped object. This is when we reveal to you that the chips we implanted in your brain are not processors, but antennas. They have been communicating with the machine hidden beneath the blanket. You aren't too pleased by the deception, but ultimately, what does it matter that the machine doing the additional processing work wasn't directly installed in your head? However, the lie runs deeper.

Beneath the blanket, as we reveal, lies a human brain. On that brain are installed duplicates of all of the antennas we installed in your brain. Furthermore, we reveal that the brain in the vat belonged to a total amnesiac and brilliant mathematician named Brian. You and Brian have, for the last two years, been doing everything together. Through your eyes, Brian has accessed the world. Given two occipital lobes, visual processing was greatly improved. Provided two prefrontal cortices, you and Brian have formed your thoughts together. The antennas work at the highest speed imaginable, so the interaction between both brains has been as quick and intimate as if we had directly installed Brian into your skull. Finally, there is no distinguishing your experiences from Brian's. All of the sensory experiences came in through the singular set of sensory organs you possess. Your lobes received that information at precisely the same time. Furthermore, visual processing (and all other processing) was shared between both brains. Your thoughts were partially formed by both brains. In some instances, a thought would originate in your brain, in others in Brian's, but in most the work appears to have been partial. At least that is what all of our measurements suggest. Indeed, you and Brian are a single subject. At least there is no reason to not think of you this way. Note further that severing your

connection to Brian will greatly affect your memories, capacities, and may even result in irreparable damage, given how your brains have adapted to communicating.

What should we take away from this? My verdict is that your subjecthood and Brian's have combined. You really are, in the sense of 'subject' we normally use, a single subject. Consider, the next time you meet someone new, when they refer to *you* (such as by calling your name or using the word 'you'), it seems they are referring to the combination of both brains. If this seems unappealing because of the spatial location of Brian, we could modify the thought experiment to somehow place Brian directly into your skull (though we'll need to go through greater trouble to tell a compelling story). Furthermore, by the definition of 'subject' I have provided, you and Brian together fit the bill. I suggest that the feeling of improved experience is *what it feels like* to combine. Or so I believe. This thought experiment is meant to prod our intuitions in the opposite direction from Goff's and Coleman's. You and Brian have combined, though there's still work to be done.

One thing I should mention now: if one wishes to deny that you and Brian are a subject, then we should also deny that our two hemispheres form a subject. Hemispherectomies can be performed where the remaining brain continues to live. Presumably, this could be done to either side of the brain. The two hemispheres utilize the corpus callosum in the same way that one communicates with Brian via the antennas. There do not appear to be any metaphysically relevant differences between the two brains and the two hemispheres. And if the subject composed of your brain and Brian's can be decomposed into two brains, and if a brain can be decomposed into two hemispheres, then it becomes clearly

arbitrary to draw the line there. There is nothing ontologically special about the specific means the two hemispheres utilize to communicate. If we can separate hemispheres, then we can separate bundles of neurons to reveal further subjects. Decomposition, at this point, has been shown to be conceivable just as much as combination has. *And that is all I need*. Interestingly, Coleman dismisses this possibility offhand when considering the irreducibility of subjects, saying we should concern ourselves with "only the relationship between ultimate-subjects and human subjects, without worrying whether the ultimate-subjects also compose subjects composing us, e.g. subjects corresponding to brain-hemispheres" (Coleman, 2013: 26).

How do we make sense of this combination, though? The way I see it, there are at least two things we can say about what is actually happening in the Brian case. The first is one I am not partial to, and that is that you and Brian remain two subjects with their own experiences. In this case, you and Brian are both having qualitatively identical experiences, but there are, at the end of the day, *two* duplicate experiences. Of course, what follows from this is that our hemispheres are *also* having two experiences. So, in the Brian case, there are four experiencing subjects with their own experiential lives. Except, as just said above, there's no reason to draw the line at the hemispheres. We have no reason to believe that the hemispheres themselves are where subjecthood bottoms out, and what I have argued in this dissertation gives good reason to believe that experience is to be found at the bottom; cutting the lines of communication between parts of the brain would simply serve to separate subjects further down. Other than at the absolute fundamental level, we will find that drawing the line between subjecthood and non-subjecthood is arbitrary, and so it must, in actuality, be trillions upon trillions of identical

experiences all the way down. If we are to hold this view of combination, then I think we should subscribe to microphenomena exhibiting Strong Character. As combination occurs, the experiences of the fundamental entities involved in that combination become more structured and less chaotic. I admit that this does appear to result in an explosion of highly vivid experience at the fundamental level of reality. To push against what might make us feel uneasy about this possibility, recall that the possession of Strong Character means that fundamental entities have an experiential vivacity of 10 out of 10. Every fundamental component of the universe, when not appropriately combined with something else, is subject to an explosion of experience. Combination, on this view, *reduces* experience, though it also imposes structure.

Why should this count as combination? If you and Brian are each having your own experience, the fact that those experiences are qualitative duplicates does not sound like combination. Combination seems to bring with it some notion of union, and here it seems as though we have escaped that. However, the reason we are even talking about combination is because the problem we are considering that panpsychism faces is the *combination* problem. My real task is to close the explanatory gap: if this is how one gets from microphenomenal experience to phenomenal experience, then we have an explanation. That being said, something *did* combine: the physical matter of each brain has a special causal relationship with the other—there is now an intimate relationship between the two. I'm happy to call this combination (in the same way that atoms can combine to form molecules). On this view, though, each micro-subject possesses its own complete experience, and these experiences are

duplicated amongst all of the members that are part of the combination. That being said, I do not champion this view, I merely believe that it can work.

I prefer a different, cleaner view. The alternative to the above is that there is only one experience being shared by both brains. I believe this is the more natural view, and I think it works best with a conception of microphenomena as exhibiting Weak Character. You and Brian are two subjects, but you share in one experience. There is no duplication; both brains are taking part in the experience. They are both processing the information, interacting, etc. I take it that this is the cleaner view, though it is also the harder one to grasp. In the previous view, it's clear what is going on. It turns out that what it means to combine is for the micro-subjects to all have the same experiential content, which does result in an explosion of experience, *but it works*. The alternative, that the trillions of subjects all partake of a singular experience, blocks experiential explosion, but we're left with the question: what does it mean to share an experience? Furthermore, what is the ontological status of 'experience' as opposed to 'subject'? I concede that I do not know. But it *is* conceivable, as I hope the Brian case shows.

Let's take the Brian case just a little further. Imagine a bunch of mini-Brians. These mini-Brians are perhaps constituted by nothing more than a few thousand neurons. We'll say we have one thousand mini-Brians. By hooking them all up together, either with the same antennas or neuronal chains or what have you, we should once again be capable of having them all *combine* so as to share an experience (or multiply it). Nothing is to stop us from breaking the mini-Brians into their composing neurons, or down further to their composing molecules, or down further to their composing atoms, or down further to the simples. But in virtue of what

do the interactions result in shared experience—which interactions result in combination? I don't know. Perhaps spatial relations are significant. Goff entertains this possibility (2017). Maybe, as I am increasingly cozying up to, microphenomenal properties really do overlap and literally merge at the fundamental level. Certainly nothing that we have said in our proposed micro panpsychist theory rules this out. In any case, one thing I hope to have made apparent: combination *is* conceivable. I have gestured at what it is like, at what it takes to achieve. Let's now return to Ub and zombies.

Remember that Ub is a combination of Red and Blue. I deny that it is any part of Red or Blue's experience that the other experience must be excluded, and I hope what I have said in this section makes clear why. Given this rejection, we can now progress on making sense of what Ub is like. Ub's experience is the experiences of its components combined. Taking what we have said above, here are two ways this might go. Perhaps Red and Blue both have a half-red/half-blue experience (or a purple one, depending on how it is that experiences blend). Alternatively, Red and Blue both share in the experience of the composite, Ub. Again, that experience can be half-red/half-blue or purple. At the end of the day, microphenomenal properties are unlikely to be red and blue, so this metaphor can be stretched only so much. It also doesn't much matter which type of composition we opt for. Whether Ub turns out to be a genuine entity in its own right, or whether our sentences about Ub's experiences turn out to be made true by the experience of Red and Blue, we can make sense of combination. On the former, Ub, being its own entity, experiences that which Red and Blue do. Of course, because of the combination, the experiences of Red and Blue modify to become half-red/half-blue or

purple. On the latter, there's just the half-red/half-blue or purple of Red and Blue themselves to account for.

Can Ub be nothing at all like Red and Blue? No. What we have said rules that out. If Ub's experience just is that of Red and Blue, then it just is Red and Blue's, either because Red and Blue are instantiating duplicates of the experience or because they are both sharing the experience. There is no room in either of these options for zombies. And so, zombies become inconceivable. Goff's pain-zombies can neither feel a billion micro-pains nor nothing at all. Not if they have combined as I have articulated above. There is no zombie option. Indeed, to propose the zombie option is just to deny that they have combined. And so we provide a board to cross the gap in explanation between the micro and the macro subjects. It is, I admit, just a board. Crossing it is precarious, and there is so much work to be done. However, the path forward is there. We need only be brave enough to cross it.

Section D: Conclusion

I presented the combination problem as an explanatory gap—the explanatory gap for panpsychism. As I have mentioned before, this isn't really fair. The original explanatory gap was an unbridgeable chasm necessitated by the appeal to brute emergence the classical physicalist believed was necessary. Our explanatory gap was but a mere crack on the sidewalk in comparison. Nonetheless, a gap in explanation it was. What the combination problem demanded was not an elaborate explanation, a bridge with every board, rail, and support beam in place. The combination problem demanded precisely what the original explanatory gap

demanded: a way forward. Had we been able to conceive of how we could move from purely physical facts to phenomenal facts, that would have been enough—that would have convinced us that progress was possible. But brute emergence disallows explanation—it's in its nature.

What I have accomplished here is the way forward we were seeking. In responding to the original explanatory gap, I reconceived fundamental physical reality. Is reality truly this way? I don't know. But if it is, then the original explanatory gap dissipates. By placing microphenomenal qualities—real, genuinely experiential qualities—at the fundamental level of reality, I made experience one of the raw materials we could appeal to, and thereby took the existence of qualia very seriously. By placing that very microphenomenal stuff as the causal basis of physical dispositions, I justified calling the theory 'physicalist'; physical dispositions are the very subject matter of physics, and those dispositions turn out to be identical with microphenomena, so it turns out that microphenomena are the referents of terms of physics. These commitments dispelled brute emergence, and with it, the original explanatory gap. The combination problem asked us how we could move from micro experiences to macro experiences. The charge, like with the original gap, was that this is impossible. Indeed, the possibility of zombies was meant to show that it is inconceivable. Yet, I tackled the intuitions underlying these charges, and I showed that we can conceive of combination. In fact, there is more than one way to make sense of it. This solves the combination problem in the same way that one would have liked the physicalist to solve the explanatory gap. It reveals that it *is* possible to move forward.

But there is much work to be done. First, what I defended in this dissertation is a conditional: if this form of panpsychism is true, then we can close the explanatory gap. However, I did not provide a thorough defense. Perhaps a nearby panpsychist theory is true instead, or perhaps panpsychism is false in all of its formulations. I doubt this latter possibility. Nonetheless, a thorough defense of panpsychism is needed. I believe, in line with Chalmers, Goff, Strawson, Skrbina, Coleman, and others that panpsychism is likely to be where the answer lies. More work to be done is in developing views on combination. The philosophical soil surrounding this problem is fertile, and I suspect clearer, stronger accounts of combination are well on their way. I am especially hopeful that intuitive accounts of combination will continue to crop up. Finally, there's a lot of taxonomical work that needs to be done. Terms such as 'experience', 'subject', 'phenomena' and its 'proto' and 'micro' varieties, etc. are still not used in a uniform manner and it isn't always clear how thinly or thickly these terms divide up their respective kinds, and having general agreement on how to use these terms is likely to dispel many illusory disagreements.

Research into panpsychism is only just starting to be taken seriously. While I did not explicitly defend the view here, it is my hope that it will soon be seen as one of the more plausible theories of mind.

Bibliography

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

Bennett, K. 2003. Why the exclusion problem seems intractable and how, just maybe, to tract it. *Noûs* 37:471-97.

Bennett, K. 2021. Why I am not a dualist. In *Oxford Studies in Philosophy of Mind Volume 1*. Oxford University Press.

Blackburn, S. W. 1990. Filling in space. *Analysis* 50:62-5.

Block, N. 1978. Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9:261-325.

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. In S. P. Stich & T. A. Warfield (eds.), *Blackwell Guide to the Philosophy of Mind*. Blackwell.

Chalmers, D. J. 2016. The combination problem for panpsychism. In G. Brüntrup & L. Jaskolla (eds.), *Panpsychism*. Oxford University Press.

Clifford, W. K. & K., C. 1878. On the nature of things-in-themselves. *Mind* 3:57-67.

Coleman, S. 2014. The real combination problem: Panpsychism, micro-Subjects, and emergence. *Erkenntnis* 79:19-44.

Coleman, S. 2015. Neuro-cosmology. In P. Coates and S. Coleman (eds.), *Phenomenal Qualities: Sense, Perception, and Consciousness*. Oxford University Press UK.

Conee, E. 1994. Phenomenal knowledge. *Australasian Journal of Philosophy* 72:136-50.

Davidson, D. 1970. Mental events. In L. Foster & J. W. Swanson (eds.). *Experience and Theory*. Clarendon Press.

Dennett, D. C. 1978. Skinner skinned. In D. C. Dennett (ed.), *Brainstorms: Philosophical Essays on Mind and Psychology*. Bradford Books.

Dennett, D. C. 2005. What robomary knows. In D. C. Dennett, *Sweet Dreams*. The MIT Press.

Fodor, J. A. 1974. Special sciences (or: The disunity of science as a working hypothesis). *Synthese* 28:97-115.

Goff, P. 2006. Experiences don't sum. *Journal of Consciousness Studies* 13:53-61.

Goff, P. 2015. Against constitutive russellian monism. In Y. Nagasawa (ed.), *Consciousness and the Physical World*. Oxford University Press.

Goff, P. 2017. *Consciousness and Fundamental Reality*. New York, USA: Oup Usa.

Goff, P. 2019. *Galileo's Error*. Pantheon Books.

Heil, J. 2004. Properties and powers. *Oxford Studies in Metaphysics* 1:223-54.

Heil, J. 2005. Dispositions. *Synthese* 144:343-56.
Heil, J. & Martin, C. B. 1998. Rules and powers. *Philosophical Perspectives* 12:283-312.
Hempel, C. 1980. The logical analysis of psychology. In N. Block (ed.), *Readings in Philosophy of Psychology*. Cambridge: Harvard University Press.
Holton, R. 1999. Dispositions all the way round. *Analysis* 59:9-14.
Horgan, T. E. 1984. Jackson on physical information and qualia. *Philosophical Quarterly* 34:147-52.
Huxley, A. 2011. The doors of perception. *Thinking Ink*.
Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-36.
Jacobs, J. D. 2011. Powerful qualities, not pure powers. *The Monist* 94:81-102.
James, W. 1890/1981. *Principles of Psychology*. Vol 1. Cambridge, MA. Harvard University Press.
Kim, J. 1984. Concepts of supervenience. *Philosophy and Phenomenological Research* 45:153-76.
Kim, J. 1987. 'Strong' and 'global' supervenience revisited. *Philosophy and Phenomenological Research* 48:315-26.
Kim, J. 1989. The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association*, 63:31-47.
Kripke, S. 1980. *Naming and Necessity*. In D. Byrne & M. Kölbel (eds.), *Philosophy*. Routledge.
Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-61.
Lewis, D. 1972. Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* 50:249-58.
Lewis, D. 1980. Mad pain and Martian pain. In N. Block (ed.), *Readings in the Philosophy of Psychology*. Harvard University Press.
Lewis, D. 1983. New work for a theory of universals. *Australasian Journal of Philosophy* 61:343-77.
Lewis, D. 1986. *On the Plurality of Worlds*. Wiley-Blackwell.
Lewis, D. 1990. What experience teaches. In W. G. Lycan (ed.), *Mind and Cognition: A Reader*. Blackwell. 499-518.
Lewis, D. 1997. Finkish dispositions. *Philosophical Quarterly* 47:143-58.
Lewis, D. 2009. Ramseyan humility. In D. Braddon-Mitchell & R. Nola (eds.), *Conceptual Analysis and Philosophical Naturalism*. MIT Press.
Maxwell, G. 1979. Rigid designators and mind-brain identity. *Minnesota Studies in the Philosophy of Science* 9:365-403.
McGinn, C. 1989. Can we solve the mind-body problem? *Mind* 98:349-66.

McKitrick, J. 2003. The bare metaphysical possibility of bare dispositions. *Philosophy and Phenomenological Research* 66:349–69.

Melnyk, A. 2003. *A Physicalist Manifesto: Thoroughly Modern Materialism*. New York: Cambridge University Press.

Mørch, H. H. 2014. *Panpsychism and Causation: A New Argument and a Solution to the Combination Problem*. PhD Dissertation. University of Oslo.

Nagel, T. 1979. Panpsychism. In *Mortal Questions*. Cambridge University Press.

Nagel, T. 1989. *The View from Nowhere*, Oxford University Press.

Peretó, J. 2005. Controversies on the origin of life. *International Microbiology: The Official Journal of the Spanish Society for Microbiology* 8:23-31.

Prior, E. W., Pargetter, R., & Jackson, F. 1982. Three theses about dispositions. *American Philosophical Quarterly* 19:251-57.

Putnam, H. 1963. Brains and behavior. In R. J. Butler (ed.), *Analytical Philosophy: Second Series*. Blackwell.

Putnam, H. 1990. The nature of mental states. In W. G. Lycan (ed.), *Mind and Cognition: A Reader*. Blackwell.

Russell, B. 1927. *The Analysis of Matter*. London: Kegan Paul.

Scharf, C. et al. 2015. A strategy for origins of life research. *Astrobiology* 15:1031-42.

Seager, W. 1995. Consciousness, information, and panpsychism. *Journal of Consciousness Studies* 2:272-88.

Seager, W. 2016. Panpsychism infusion. In Bruntrüp & Jaskolla (eds.), *Panpsychism: Contemporary Perspectives*. Oxford University Press.

Shapiro, L. A. 2010. Lessons from causal exclusion. *Philosophy and Phenomenological Research* 81:594-604.

Shoemaker, S. 1982. The inverted spectrum. *Journal of Philosophy* 79:357-81.

Skrbina, D. 2007. *Panpsychism in the West*. Bradford.

Smart, J. J. C. 1959. Sensations and brain processes. *Philosophical Review* 68:141-56.

Strawson, G. 2006. Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies* 13:3-31.

Strong, C. A. 1919. *The Origin of Consciousness: An Attempt to Conceive the Mind as a Product of Evolution.* Macmillan and Company.

Yablo, S. 1987. Identity, essence, and indiscernibility. *The Journal of Philosophy* 84:293-314.