PHYSICALISM AND PHENOMENAL CONSCIOUSNESS

by

Adam Curran Reid

A thesis submitted to the
Department of Philosophy
in conformity with the requirements for
the degree of Master of Arts

Queen’s University
Kingston, Ontario, Canada
August 2010

Copyright © Adam Curran Reid, 2010
To Tina Vashistha: I could never have come this far without your love and support. For your strength, your inspiration, and for always believing in me, even when I doubted myself, I dedicate this to you.

xo, always - Adam
Abstract

The following thesis is concerned with the contemporary debate in the philosophy of mind concerning the nature of phenomenal consciousness (viz. subjective experience, or *qualia*). My primary aim is to adjudicate the ongoing dialectic between dualists and physicalists regarding the ontological status of phenomenal consciousness—physical or nonphysical—by examining the two major arguments most commonly deployed *against* physicalism: namely, the zombie argument (Chalmers), and the knowledge argument (Jackson). I conclude by showing that once physicalism has been shorn of the various doctrines that it need not and ought not accept—that is, once we are clear about what, precisely, the fundamental doctrine of physicalism actually *is*—it becomes clear that these arguments do not go through, and that dualism has not made its case. I also argue that the task of actually disarming these arguments (*in the right way*) is critically instructive to contemporary physicalists, as this helps to nourish a clearer overall understanding of what physicalism (properly understood) is, and is not, committed to.

In Chapter One I lay the groundwork for the aforementioned anti-physicalist arguments by explaining precisely what is meant by the phrase “phenomenal consciousness” and its various synonyms. I then briefly summarize the mind-body problem and articulate the so-called *explanatory gap* therein. Chapter Two looks at the zombie argument (as articulated by David Chalmers, 1996) and finds that the argument itself, as stated, actually has very little to do with defending dualism against physicalism, but rather is ultimately an argument for epiphenomenalism—which, I argue, is untenable. Chapter Three examines Frank Jackson’s knowledge argument against physicalism. Here I show why the argument itself fails to support property dualism, but also why the standard physicalist objections to it fail. I argue that Mary does indeed learn facts *about the world* upon her release, and that physicalists must face up to this squarely. I then show that physicalism (properly understood) is entirely compatible with this admission. Chapter Four examines a kind of rehabilitated version of the zombie argument in the context of a larger discussion about the relation between conceivability and possibility.
Acknowledgements

First and foremost I would like to thank the Department of Philosophy at Queen’s University, for its generous financial assistance during the year, and for its commitment to providing a friendly and intellectually supportive atmosphere in which to pursue graduate studies in philosophy. Looking back on the experience, the past year has provided everything I had hoped for from a Master’s degree: lasting friendships; strong philosophical instruction and lively debate; and, above all, a solid foundation and hunger for more study.

Thanks also to the Ontario Ministry of Training, Colleges and Universities, whose financial support in the form of an Ontario Graduate Scholarship helped to make this year possible for me.

Special thanks to my supervisor, Deborah Knight, whose enthusiastic comments and helpful criticisms of earlier drafts of this manuscript were of great assistance. As a supervisor, Deborah struck exactly the right balance between gentle guidance and firm nudging, and I’m convinced that my argument is clearer, more persuasive, and more original because of this. Thanks also to Deborah for first introducing me to the philosophy of mind in general, and to the ‘hard problem’ in particular. It was her engaging lectures on the subject that first ignited my philosophical curiosity for these questions as a senior undergraduate student, so it’s fitting that our many discussions throughout the course of this project have served only to deepen my interest, and to kindle my enthusiasm all the more.

I would also like to thank the Queen’s Philosophy graduate student community. Portions of this thesis were presented as works-in-progress at the Graduate Colloquium, and I am grateful to Katherine Wayne for organizing this series, and to all who participated in making it a success. In particular, the comments and suggestions raised here were invaluable in showing me where the structure and development of my argument could be made clearer or more direct.

I must also give special thanks to my friends and family for all their love, patience, and support throughout the past year. In particular I would like to thank Catherine French, Torin Doppelt, and especially Tina Vashistha — all of whom have generously participated in more late-night philosophical conversations about qualia, zombies, and imaginary super scientists named Mary than any sane person should have to endure. To be sure, virtually nothing worth reading in contemporary philosophy is ever produced entirely in isolation, and this thesis is no exception.
Finally, as with any undertaking of this size and scale, this project owes a singular debt to the various philosophers whose ideas I discuss — in particular David J. Chalmers, whose absorbing book *The Conscious Mind* is the touchstone work that this thesis explores. Though the final argument defended here is fundamentally critical of Chalmers’ overall account, the exceptional clarity and rigor with which he consistently expresses and defends his position were both instructive and inspiring in my efforts to formulate an alternative stance on the consciousness debate. Indeed, as someone who, in recent years, actually thought himself to be an aspiring political philosopher, it was the experience of reading *The Conscious Mind*, more than anything else, that taught me what a truly fascinating field of study the philosophy of mind can be.
# Table of Contents

Abstract i

Acknowledgements ii

Chapter One: Experience and the Explanatory Gap 1

Chapter Two: The Zombie Argument and its Epiphenomenal Premise 9

2.1 Introduction: structure and goals of the chapter 9

2.2 Zombies, functionalism, and the hard problem of consciousness 10

2.3 Thomas on the incoherence of zombie claims to phenomenal self-knowledge 14

2.4 The phenomenology of experiential belief: in defence of Perry’s neo-Humean “epistemology of experience” 20

2.5 Conclusion 30

Chapter Three: Why Mary Couldn’t Know all the Facts 33

3.1 Taking stock and the move to antecedent physicalism 33

3.2 Engaging the knowledge argument 39

3.3 Physicalist strategies for disarming the knowledge argument 43

3.4 On knowing what it’s like: acquaintance plus recognition 49

3.5 Flanagan on the distinction between linguistic and metaphysical physicalism 55

Chapter Four: Conclusion 58

4.1 In defence of antecedent physicalism 58

4.2 Refrain: zombies, conceivability, and possibility 63

Bibliography 70
CHAPTER ONE

Experience and the Explanatory Gap

In his classic paper “What is it Like to be a Bat?” Thomas Nagel provides the following widely employed analytic definition of phenomenal consciousness.

No doubt it occurs in countless forms unimaginable to us, on other planets in other solar systems throughout the universe. But no matter how the form may vary, the fact that an organism has conscious experience at all means, basically, that there is something it is like to be that organism. There may be further implications about the form of the experience; there may even (though I doubt it) be implications about the behaviour of the organism. But fundamentally an organism has conscious mental states if and only if there is something it is like to be that organism—something it is like for the organism. We may call this the subjective character of experience.¹

Excluding any potential misgivings about the philosophical merits of such armchair theorizing across species, few deny that Nagel’s definition helpfully describes a phenomenon that actually exists in the world—even if we can only really be sure that it exists in the case of human beings.² However, despite such consensus at the purely descriptive level, there is much debate in contemporary philosophy of mind surrounding the ontological status of


²I confess, to my mind restricting the special scope of phenomenal consciousness to human beings alone seems little more than bald anthropocentrism. This particular debate, however, falls outside the scope of this treatise, and I will not consider it here.
phenomenal consciousness. This is because Nagel’s definition—namely, that a mental state is a conscious state just in case there is “something it is like” to be in that state—is, by itself, “topic-neutral” with respect to the further question of whether the nature of the phenomenon described is, ontologically, physical or nonphysical in kind.

On one end of this debate are those who insist, in varying degrees and with various arguments, that it is just obvious that consciousness is irreducibly ontologically distinct in kind from, say, tables and chairs, since what it means for something to be physical is, fundamentally, for it to be spatially extended. Proponents of this view, called “dualism,” typically preface their arguments by highlighting the following sorts of considerations. Consider the brain. The average human brain weighs 14oz, is dull grey in hue and is probably somewhat soggy to the touch. More to the point, each of these attributes clearly betokens the fact that the brain is an extended and hence a physical thing. But what about the mind? Quite unlike the brain, the mind is something that is defined in terms of such seemingly intangible phenomena as thought, belief, perception, memory, and so on. Now, on the face of it, it seems pretty clear that all of these phenomena are not the sorts of things that one can plausibly describe as extended. It would be absurd, for example, to wonder how much a given thought or memory “weighs.” Likewise, even though there is no doubt some very specific neurophysiological event(s) taking place inside my head when I experience, say, a green afterimage, there is certainly no corresponding “green spot” on my brain—rather, there is just so much whirring microscopic electro-physiochemical activity. But where, then, does the afterimage occur?

The reason I draw the reader’s attention to these remarks is to underscore the considerable raw intuitive force of the classic Cartesian stance on the mind-body problem, which holds that while the brain is a material thing that exists in both time and space, the mind is immaterial in nature, and thus exists in time but not space. To be sure, dualism has come a long way since the time of Descartes; but it is arguably these same sorts of intuitive reflections that ultimately undergird even its more sophisticated contemporary forms.
The clearest way to appreciate just what is at stake in the contemporary debate about the nature of consciousness is to consider the related question of what sort of account, in principle, would even count as properly explanatory in the first place. More to the point, the real issue here is whether the essentially empirical methodology of science is an appropriate touchstone in the task of constructing a fundamental theory of consciousness. Dualists, broadly speaking, are inclined to reject this suggestion as fundamentally misguided. The spirit, if not quite the letter, of dualist scepticism on this point can arguably be traced back to an ingenious and widely influential thought experiment from G.W. Leibniz. Here is the relevant passage.

Moreover, it must be confessed that perception and that which depends upon it are inexplicable on mechanical grounds, that is to say, by means of figures and motions. And supposing there were a machine, so constructed as to think, feel, and have perception, it might be conceived as increased in size, while keeping the same proportions, so that one might go into it as into a mill. That being so, we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception. (Leibniz, Monadology, 1714, para. 17 [Latta translation])

The reason for the enduring relevance of this idea in contemporary philosophy of mind is clear enough. If you are a dualist, then the upshot of taking seriously the moral of Leibniz’s “Mill argument” is just to underscore the fact that we are, and perhaps always will be, in precisely the same position with respect to our knowledge of the brain, such that no amount of scientific observation—no matter how sophisticated our investigative techniques become, or how exhaustive we are in mapping the precise microphysical operations that occur therein—could, in principle, even begin to tell us how, as Colin McGinn puts it, “Technicolor phenomenology [could] arise from soggy grey matter.”3 The following excerpt from McGinn’s seminal essay on the subject will serve nicely as a contemporary update on the Leibniz intuition, as well as a concise summary of the fundamental problem this thesis seeks to address.

How could the aggregation of millions of individually insentient neurons generate subjective awareness? We know that brains are the de facto causal basis of consciousness, but we have, it seems, no understanding whatever of how this can be so. It strikes us as miraculous, eerie, even faintly comic. Somehow, we feel, the water of the physical brain is turned into the wine of consciousness, but we draw a total blank on the nature of this conversion.4

The fundamental nub, then, or “hard core” of the mind-body problem is the persistence and recalcitrance of an apparent “explanatory gap” (a term due to Levine5) that exists between the empirically detectable goings-on in our brains, and our total mystification as to how (much less why) any of these goings-on should somehow give rise to a subjective inner life—the latter, according to dualists, being a phenomenon that is in principle empirically undetectable. In a landmark paper, David Chalmers has famously coined this “the hard problem of consciousness.”6 Indeed, as we will see, Chalmers is arguably the foremost contemporary defender of dualism working in the philosophy of mind today. Interestingly, though Chalmers provides a variety of sophisticated analytic arguments to show that there are principled, systematic reasons for thinking that the explanatory gap between the physical and the phenomenal is indeed scientifically intractable (at least by means of conventional reductive analyses), Chalmers also suggests that the “gap,” though every bit as deep as the Leibniz example implies, is actually much more narrow than many dualists have traditionally supposed. The reason for this, he argues, is because the hard problem of consciousness refers only to the phenomenon of subjective experience. Thus, in contrast to what is implied by Leibniz’s phrasing, Chalmers argues that nearly all mental phenomena—e.g. belief, memory, report, perception, control, and so on—are straightforwardly scientifically tractable, and therefore fall into a category that he terms “the easy problems of consciousness.” “Easy,” of course, is a relative term, and indeed Chalmers is careful to emphasize that he in no way means to suggest that the “easy problems” are not

4Ibid. p.369.
still difficult problems in their own right; surely they are. The crucial point is that unlike the hard problem, these are all the sorts of phenomena that are in principle fully amenable to conventional reductive methods of analysis, even though it will surely take many more years until these methods are perfected, and hence, until such analyses are complete.

The central insight behind Chalmers’ easy problems/hard problem distinction is that with the easy problems, ultimately, all that needs explaining are various structures, functions and abilities. More precisely, all that is conceptually required for an exhaustive explanatory account of these phenomena is that we are able to show how a particular cognitive system is able to perform the associated functions. Once this is done, the explanatory burden has been fully satisfied. Chalmers’ fundamental claim, then, is that most aspects of our mental life are in principle amenable to functionalist analyses, since the particular phenomena that cry out for explanation here, in the end, just are functions and abilities. However, as Chalmers rightly stresses, there is evidently a good deal more to the life of the mind than the mere performance of various functions and the regulation of behaviour. There is also the further phenomenon of qualitative experience itself—i.e. the manifest what-it-is-like aspects or “raw feels” that accompany, for example, red-seeing, chocolate-tasting, pain-feeling, and so on. The problem, however, is that just as evident as the fact that these phenomenal properties—these qualia, as the jargon has it—really do exist, it seems equally clear that their occurrence is not something that we can analyze in terms of structures, functions and abilities, since it is always possible to give an explanatorily exhaustive account of what the mind does in purely functional terms that neither invokes nor implies the presence of subjective experience. Thus, in the end, it is the brute existence of subjective experience itself that is the core explanandum of the hard problem of consciousness. Chalmers explains,

With experience...physical explanation of the functions is not in question. The key is instead the conceptual point that the explanation of functions does not suffice for the explanation of experience. This basic conceptual point is not something that further

---

7Ibid. See esp. pp. 382-83.
8This, in essence, is the central thesis that I try to refute in Chapter Two.
neuroscientific investigation will affect.\textsuperscript{9}

Chalmers has suggested that at the center of this puzzle what we are really facing are two fundamentally distinct concepts of mind; concepts whose relationship, of course, we would like to make clear. On the one hand, there is the “psychological concept of mind,” where the mind is characterized by what it \textit{does}; and on the other hand there is the “phenomenal concept of mind,” where the mind is characterized by the way it \textit{feels}.\textsuperscript{10} Nor is this a distinction that we can simply dissolve or explain away as a merely intuitive chimera in our “pre-theoretical” thinking about the mind. Indeed, if Chalmers is right, then the distinction is \textit{analytic}, because any attempt to phenomenologically describe the particular qualitative character of how something feels or seems is \textit{conceptually independent} of any explanation that we might give of the underlying structures, functions and abilities in question. In sum, reductive analysis—i.e. the attempt to explain a given high-level phenomenon wholly in terms of more basic physical processes—cuts no ice in answering the fundamental explanandum of how and why subjective experience \textit{arises} from the brain.\textsuperscript{11}

Though they are importantly distinct in various, often technical respects, the preceding arguments I have just sketched from Leibniz, McGinn, and Chalmers are nevertheless unified by a fundamental common denominator, a shared sentiment that can be described, following Loar, as the “anti-physicalist intuition.”\textsuperscript{12} More precisely, the anti-physicalist intuition holds that phenomenally conscious mental states, or \textit{qualia}, are irreducibly ontologically distinct from physical mechanisms in the brain. Notice, however, the implicit form that each argument takes in trying to give shape to this core intuition. Consider: each argument begins by vividly calling our attention to the considerable difficulties we immediately face in trying to understand \textit{how it could be} that phenomenal consciousness arises from the brain.

As McGinn observes, “The mind-body problem is the problem of understanding how the 
miracle is wrought, thus removing the sense of deep mystery. We want to take the magic out 
of the link between consciousness and the brain.” And yet, as the Leibniz example tries to 
show, there is at least a strong prima facie case for granting that there may be conceptual 
reasons for conceding that the actual form that such understanding would necessarily have 
to take could not be grounded in purely “mechanical” (i.e. functional, physiochemical) terms 
of low-level neurological processes and events in the brain. Again, the problem with this 
sort of theoretical undertaking is that it would serve merely to underscore the explanatory 
gap between brain states and their phenomenal correlates. McGinn agrees: “The brain 
has physical properties we can grasp, and variations in these correlate with changes in 
consciousness, but we cannot draw the veil that conceals the manner of their correlation.”

Chalmers, in effect, argues that the Leibniz conjecture illustrates what is actually a 
fundamental analytic point about the inherent limitations of reductive analysis as such. 
Chalmers, however—and, indeed, Frank Jackson, who we will discuss at length in Chapter 
Three—goes beyond both Leibniz and McGinn when he argues that the objective truth of 
physicalism stands or falls with the success or failure of explanatory reduction. One thing 
to keep in mind, then, as we move to consider the two major arguments against physicalism 
in Chapters Two and Three is whether or not this particular tenet of contemporary dualism 
is philosophically well-founded. I will argue that it is not. There is, rather, an alternative 
and, indeed, an entirely more plausible way of construing the (possible) objective truth of 
physicalism—a strategy that, if embraced, would enable physicalists to sidestep many of the 
standard criticisms that dualists have raised. Though we are not yet in a position to state 
this position fully, the following will serve as a rough sketch of the general idea. In short, 
I will argue that physicalists should concede that explanatory reduction of phenomenal

---

14 Ibid. p. 378. 
15 See esp. Section 7. (pp.121-122) of Chapter 3 of The Conscious Mind, called “Can Consciousness be 
Reductively Explained?” pp. 93-122. See also Section 5 of “Facing up to the Problem of Consciousness,” 
consciousness may, in the end, prove conceptually impossible, but they should deny that this alone furnishes a decisive justification for our moving, inferentially, from what is an essentially epistemological claim to the strong metaphysical conclusion that the world must therefore include ontologically nonphysical phenomena.

Thus, on one level, the account that I will provide here is (at least, broadly speaking) consistent with a so-called “mysterian” stance on the mind-body problem. That said, part of my agenda throughout will also be to consider which side of the consciousness debate—dualism or physicalism—properly bears the burden of proof in regards to the metaphysical question, and thus which side constitutes the appropriate default in the ongoing dialectic. My central thesis will be that once physicalism has been shorn of the various doctrines that it need not and ought not accept—that is, once we have gotten clear about what, precisely, the fundamental doctrine of physicalism actually is—it will have become clear that the various anti-physicalist arguments that dualists have provided do not go through, even though the task of actually disarming these arguments (in the right way) is, often crucially, critically instructive to contemporary physicalists in better articulating their own position.

CHAPTER TWO

The Zombie Argument and its Epiphenomenal Premise

2.1 Introduction: structure and goals of the chapter

The purpose of this chapter is to engage critically with David Chalmers’ modal argument against physicalism: namely, the proposed logical conceivability of zombies. More specifically, I wish to explore (and, ultimately, to exploit) what appears to be a singularly difficult conceptual problem faced by sponsors of the zombie argument, an implication that Chalmers has termed “the paradox of phenomenal judgment.” Ultimately, my goal is deflationary, and can be described as pursuing two principal goals, in roughly the following order. First, following Nigel Thomas’s instructive lead, I will suggest that the concept of philosophical zombies—or, more broadly, that of a “zombie possible world”—is revealed as not expressing

\textsuperscript{17} Chalmers, \textit{The Conscious Mind}, p. 181.
a coherent possibility. Next, following a critique of Chalmers advanced by John Perry\textsuperscript{19}, I will argue that the price of conceding that zombies are not logically conceivable is actually relatively modest, as it demands only that one must reject epiphenomenalism about consciousness.\textsuperscript{20} Certainly, it does not imply that one \textit{must} also embrace a physicalist, much less an eliminativist, stance on the ontology of (qualitative) mental content. Indeed, as we shall see, it does not follow from merely demonstrating that consciousness is causally efficacious—and hence conceptually necessary for the full functional explanation of our behaviour—that one has done anything to undermine (or even to address) Chalmers’ overall non-reductive stance on the metaphysical nature of phenomenal consciousness. The fundamental ontological debate between physicalists such as Perry and dualists such as Chalmers, rather, engages with a separate, indeed a deeper question than that of epiphenomenalism; namely, (a) can conscious experience be reductively explained?, and (b) what follows from (a)? The zombie argument, we will see, turns out to be quite unhelpful when it comes to providing any unique insight into this further question. In fact, we may even come to see that, on the whole, “philosophical zombism” is actually responsible for generating more methodological confusion than clarity regarding how both sides of this debate ought best to engage with each other. I turn now to a brief overview of the structure and putative purpose of the zombie argument.

2.2 Zombies, functionalism, and the hard problem of consciousness

A philosophical zombie is a hypothetical being that is functionally, behaviorally, and physically identical to a normal human being, but who lacks any conscious experience. To help


\textsuperscript{20}This does, of course, create problems in itself: by rejecting epiphenomenalism it seems the property dualist is also committed to rejecting the premise that the physical domain is causally closed. In this way, the property dualist becomes faced with the further task of having to give a plausible account of behavioural causation in mental-physical \textit{interactionist} terms.
flesh this idea out, Chalmers invites us to imagine a “zombie possible world,” by which is meant a logically conceivable hypothetical world that exists parallel to our own, and which is externally *indistinguishable* from our world. To do this, we are to imagine ourselves (along with everyone else) as having a “zombie twin”—a doppelganger—who says, does, in short, who lives *exactly* the same life that we do in *every* sense that is externally observable from the third-person perspective, as well as from the objective standpoint of a perfectly completed physics of mind. The crucial difference, Chalmers suggests, is that whereas the “sum” of my life will necessarily involve an irreducible “qualitative” dimension—namely, the internal, non-publicly observable phenomena of *subjective* “raw feels” (i.e. *qualia*) that I experience within my own private mental interior—the sum of my zombie twin’s “life” can, in principle, be *fully* accounted for (i.e. without remainder) in purely functional terms of the complex microphysical operations that make up, indeed, that exhaust his internal cognitive system. The reason for this, of course, is that, by definition, there is *nothing it is like* to be a zombie and hence there simply are no *further* phenomena here that need to be explained. By contrast, the brute fact that the events of my life, unlike those of my zombie twin, are accompanied by a distinct phenomenal *feel* is, we are told, itself sufficient grounds to establish that a purely functional account of my life would be seriously incomplete.

Chalmers argues persuasively that the central problem with reductive theories is their failure to regard subjective experience itself as an *explanandum* in its own right. The problem, as he sees it, is that any reductive account of the structure, functions, and complex interaction of brain processes just will not suffice—indeed, could not suffice—to explain why any of the capacities and abilities that these various processes produce in us should also give rise to a subjective inner life. Stated differently, there does not appear to be any necessary *a priori entailment relation*, in the strong conceptual sense, between our ability, as adaptive, self-aware organisms, to perform various behavioral acts and feats of information processing, and the (apparently passive) further phenomenon of our also being conscious *experiencers*. Why, in other words, does the entire complex informational processing that occurs in our
brains, and which, we are told, ultimately governs our behaviour, not simply take place “in the dark?” Chalmers has termed the “hard problem of consciousness,” and, indeed, a key part of the motive behind the zombie argument is surely to press home this very point—i.e. that to accept that zombies are logically conceivable just is to accept, inter alia, that the preceding question, at the very least, expresses a coherent possibility.

Now, the above summary is intended only as a rough sketch of the basic structure and general strategy of the zombie argument. The intricacies and, I will suggest, the internal problems with the argument will be brought out in the course of the critique that I will now mount. Nor should this summary be taken as an attempt to raise, much less to press, the altogether speculative claim that Chalmers’ actual stance on the zombie question is really the strong claim that one demonstrably does not take seriously the hard problem of consciousness, and the “explanatory gap” therein, unless one also accepts the logical possibility of zombies. Indeed, this would be a grave misrepresentation and quite unjust oversimplification of his view. Moreover, this would be simply to ignore the fact that Chalmers has, in various places, offered a variety of sophisticated arguments for his non-reductive position—many of which, I believe, are far more compelling than the zombie scenario—which make no appeal to zombies whatsoever. The point is that one can argue, both directly and indirectly, for the hard problem-easy problems cum “two concepts of mind” distinction (and, ultimately, for the property dualism that this supports) quite independently of any recourse to the zombie world thought experimental methodology. It is, however, fair to say that Chalmers clearly believes that the zombie world scenario helps to make vivid, and perhaps even to unify, these other zombie-neutral approaches, thus lending additional support to his overall non-reductive view. For the remainder of this chapter, I will argue that the opposite is in fact the case: that despite its prima facie plausibility, the zombie argument, as stated, is actually doing more harm than good.

22 See esp. his paper, “Facing Up to the Problem of Consciousness,” Idem. See also Chapters Three and Four of The Conscious Mind, entitled (Three) “Can Consciousness be Reductively Explained?” and (Four) “Naturalistic Dualism.”
There remains one further feature of the zombie argument that must be brought to light before we can examine critically the internal coherence of the concept itself. This, of course, is the necessary, built-in premise that the physical world is a causally closed system. Consider: if (a) my zombie twin says and does *everything* that I say and do (i.e. by all external accounts, we are living *identical* lives) and if (b) he lacks any and all conscious experiences, then it follows that *my* subjective experiences, and the phenomenal mental properties of which they are composed, are causally inert. The reason for this is that the same reductive-functionalist account that we would employ to describe the complex microphysical processes that, ultimately, realize the actions of my zombie twin can also be applied, with *exactly equal explanatory force*, to the task of giving a complete third-personal, causal understanding of *my* actions, because, after all, my zombie twin and I are, *ex hypothesi*, behaviorally indistinguishable beings. The upshot, then, is that by accepting the coherence of the possibility expressed by the behaviorally indistinguishable stipulation, we thereby also commit ourselves to saying that the fact of phenomenal consciousness in our world is *explanatorily irrelevant* to the task of constructing a complete causal understanding of our behavioral functioning. Qualia, on this view, are simply along for the ride. They are the epiphenomenal exhaust, or gratuity, so to speak, of the complex causal operations that occur in our physical brains and bodies. Though we will not engage with the epiphenomenal premise directly until the final section of this chapter, it will be helpful to bear this in mind as we move to explore Thomas’s reductio critique of the zombie argument, and, more importantly, as we assess the plausibility of Chalmers’ response. I turn to this task now.
2.3 Thomas on the incoherence of zombic claims to phenomenal self-knowledge ²³

Nigel Thomas’s critique of the zombie world scenario is based in what he suggests are the ultimately self-defeating implications that follow from taking seriously the full demandingness of the behaviorally indistinguishable stipulation. For the sake of economy, I will occasionally refer to this as the “B.I. clause.” Though what follows is by no means the only problematic implication that arises for zombie advocates who accept this stipulation (as we shall see in the next section), the particular objection that Thomas highlights concerns the conceptual problems that ensue once we reflect carefully on the fact that zombies would claim to be conscious. For illustrative purposes, we are asked to first assume the role of one who seeks to defend the view that zombies are, in fact, logically conceivable. With this in mind, Thomas begins his analysis with the uncontroversial observation that there are three (and, indeed, only three) positions available to us in our attempt to parse this particular utterance from our zombie twin. These are: (a) that the claim is false; (b) that the claim is true; or (c) that the claim is meaningless. From here, Thomas shows that whichever line we might take on this point, we inevitably lapse into incoherence and contradiction in our attempts to articulate a response.

Consider, first, response strategy (a): that our zombie twin’s claim is false. Here is the problem with this approach. Notwithstanding the fact that I appear already to “know” that my zombie twin is “mistaken”²⁴—consciousness, after all, simply does not obtain in

²³The first half of this section (roughly, pp. 14-17.) is essentially just a critical summary of Thomas’s central argument, and should not be read as original argumentation. For a more thorough presentation of the argument, see his paper “Zombie Killer,” Idem.

²⁴Note that the proposed falsity of his utterance could not be the product of his lying. Consider: given (a) that I do not have such an intention to purposefully misrepresent when I claim to be conscious, we know also (b), that because whatever internal articulation mechanism of, let us say, “epistemic insincerity” that would necessarily engage in the event that I were to form, and to act from, the intention to lie does not occur in me (because, again, I take myself to be advancing a true proposition), then (c) neither does it occur in my zombie twin. In short, speech is a behavioral act; hence the B.I. clause, coupled with the physically indiscernible stipulation—or P.I.—ensures that zombie-Adam, who, after all, is my cognitive
his world—I cannot ignore the fact that whatever cognitive articulation mechanism it is (and there must be some such mechanism) that enables me to go from the fact of my consciousness to a (true) belief in my own consciousness is, *ex hypothesi*, the very same cognitive mechanism, acting in exactly the same way, that enables zombie-Adam to go from the fact of his non-consciousness to the equivalent (mistaken) belief that he is conscious. If, however, the two cognitive mechanisms in question are physically identical, functioning in precisely the same way, then it would appear that I am not, strictly speaking, entitled to the presumption that my own claim to consciousness is any more secure, much less truth-bearing, than that of my zombie twin. The problem, then, with response strategy (a) is that, by granting that mistaken zombies are logically possible, we unavoidably throw strong doubt on the epistemic reliability of our own phenomenal judgments. In other words, response strategy (a) demands that we also accept the real possibility that we might be zombies.

Now let us consider response strategy (b): that zombie-Adam’s claim is *true*. The problem here is closely related to the one we just considered, in regards to response strategy (a). Response strategy (b) goes as follows. Given the apparent reductio that we end up with in purporting to describe zombie-Adam’s claim as false, perhaps we might say instead that his claim is “true”, but true in a *different sense* than normal-Adam’s claim is true. The idea here is that when zombie-Adam claims that he is conscious, he actually means something different from what normal-Adam means when he advances the equivalent claim. Very roughly, the distinction is as follows: when *I* claim to be conscious, the complex phenomenon that I refer to involves both of the following: namely, (a) my *performative capacity* to engage in all the functions and abilities implied by what Chalmers has termed the “psychological concept of consciousness” (so, roughly speaking, “responsiveness to my environment”) as well as (b) my *introspective experiential acquaintance* with the various subjective characters (i.e. the *qualia*) that normally accompany my behavioral functioning.
(i.e. what Chalmers calls the “phenomenal concept of consciousness”). Conversely, my zombie twin refers only to (a)—to the psychological concept of consciousness—when he claims to be conscious. In the zombie literature, this distinction is sometimes referred to as that between “consciousness” (or, perhaps better, “full consciousness proper”) and what Moody has dubbed “consciousnessz”.26

Despite some initial plausibility, however, this strategy fails to the extent that it is straightforwardly inconsistent with the full demandingness of the B.I. clause. The incoherence of this strategy can be stated very simply. Consider: the very fact that I am able to conceive and articulate the (putative) distinction between consciousness and consciousnessz implies, ex hypothesi, that my zombie-twin is able to make this same distinction. Thus, insofar as I claim to be conscious in the full sense—invoking both the psychological and the phenomenal concepts of mind—then so does my zombie twin. To deny this implication would be equivalent to denying the B.I. clause itself, which would in turn be tantamount to denying the logical conceivability of philosophical zombies. Chalmers provides an especially vivid account of what’s wrong with this strategy.

To strengthen the sense of paradox, note that my zombie-twin is himself engaging in reasoning just like this. He has been known to lament the fate of his zombie twin, who spends all his time worrying about consciousness despite the fact that he has none. He worries about what that must say about the explanatory irrelevance of consciousness in his own universe. Still he remains utterly confident that consciousness exists and cannot be reductively explained.27

So much for response strategy (b). Finally, we come to response strategy (c): that zombie-Adam’s claim is meaningless. Very roughly, the suggestion being proposed here is that my zombie-twin’s claim could not be meaningful because “consciousness,” in his universe, is what philosophers call a “non-referring term.” However, as Thomas quite rightly points out, “mere non-referring terms will not get us meaningfulness, at least not the right

---

kind."\textsuperscript{28} To illustrate this point, Thomas offers a helpful example of a meaningful expression whose subject is a non-referring term: suppose, he suggests, someone were to say to you “I have a jabberwock in my pocket.”\textsuperscript{29} Thomas argues convincingly that the mere non-existence of jabberwocks is insufficient to render this assertion meaningless. Assuming that one is familiar with Carroll’s \textit{Alice in Wonderland} (and hence, with the fictional creatures “jabberwocks”), one also understands what the truth conditions of this utterance would have to be; that is, (a) there would have to be at least one jabberwock in existence, and (b) our interlocutor would, \textit{in fact}, have to have one in his pocket. Now, in view of the fact that, metaphysically speaking, neither condition obtains in our world, the right conclusion to draw is that the assertion is, simply, false. The possibility expressed by the assertion itself, however, is quite coherent, and hence was perfectly meaningful all along. If we accept this story (and I believe there is a lot to be said for doing so), then we will naturally follow Thomas in concluding that our zombie-twin’s claim to full consciousness proper is clearly meaningful, since it plainly expresses a conceptually coherent, veridically analyzable logical possibility. Thus, the right conclusion to draw from our zombie twin’s assertion is not that it is meaningless, but rather that it is false, since, \textit{ex hypothesi}, consciousness does not exist in his universe.

So much for response strategy (c). Where, then, does this leave us? We appear to have come full circle and are thrown back, as it were, into response strategy (a). But more importantly, we seem to have arrived, inexorably, at the following binary choice. Either we accept that mistaken zombies are a logical possibility \textit{along with} the corollary that \textit{we} might actually be zombies; or we simply reject, as incoherent, the zombie concept itself, thus preserving the epistemic reliability—indeed, the verity of our own phenomenal judgments about consciousness. Thomas suggests that we elect the latter choice, and I agree. Chalmers, on the other hand—though he agrees that response strategy (a) is the only plausible reading of a zombie’s claim to consciousness—advances an alternative proposal,

\textsuperscript{28}Thomas, “Zombie Killer,” p. 175.
\textsuperscript{29}Ibid.
a proposal that he suggests permits the zombie advocate to sidestep the unhappy binary choice outlined above. In short, Chalmers denies that by accepting the logical possibility of mistaken zombies, we necessarily degrade the integrity of our own phenomenal judgments about consciousness. Very roughly, the response that Chalmers offers proceeds as follows. The paradox of our asserting what looks, at first blush, to amount to an untenable double standard about truth between two exactly equivalent claims about consciousness (i.e. mine, and that of my zombie twin) can be diffused if we accept that the justification for my belief, contra that of my zombie twin, derives not only from the cognitive mechanism(s) that enable me to form and articulate my belief, but also from my direct evidence—evidence which, of course, my zombie twin lacks (consciousness, after all, does not exist in his world). Chalmers also observes that, strictly speaking, zombies would not have the same beliefs as we do, arguing that conscious experience is itself partially constitutive of the (phenomenal) content of our beliefs. Again, given that zombies, ex hypothesi, have no conscious experiences, the contents of their beliefs about consciousness will differ crucially from ours. Hence, Chalmers concludes, the grounds for our own phenomenal beliefs are not fatally undermined by granting the logical possibility of mistaken zombies.

I do not think that this response works. The problem is that it can be seen to be straightforwardly vulnerable to the same objection that Thomas raised (and which Chalmers concedes) against response strategy (b), that our zombie twin’s claim is true. The force of this objection can actually be brought out nicely by simply extending the paradoxical scenario that Chalmers himself described so well in the brief passage we looked at above. Consider: facing up to the full implications of the extreme demandingness of the B.I. clause, Chalmers would be forced to concede that an exactly equivalent argument to the one he has just given in response to Thomas’s ultimatum would also be advanced, verbatim, by zombie-Chalmers in response to zombie-Thomas in the course of their zombic debate. In other words, zombie-Chalmers—who, we are told, is “utterly confident” (though “monumentally...

---

30For Chalmers’ more detailed account of the argument, see Chapter 5 of The Conscious Mind, called “The Paradox of Phenomenal Judgement.” See esp. pp. 196-98.
deluded,”31) that he is conscious (and, recall, is behaviorally indistinguishable from actual-Chalmers)—would equally cite the “fact” of his direct evidence of phenomenal consciousness as supplying him with the crucial justificatory edge over his zombie twin in asserting, as well-founded, the reliability of his own beliefs.

And, indeed, the story does not stop there; for again, taking the B.I. clause seriously, Chalmers must also concede that his zombie twin—let’s call him “zombie-Chalmers Alpha”—likewise concedes that his zombie twin—call him “zombie-Chalmers Beta”—would argue, in precisely the same way, for the greater justificatory support for his beliefs contra those of his zombie twin—call him “zombie-Chalmers Gamma”—and so on, and so on, ad infinitum. Hence it would appear that the initial problem with response strategy (a)—that our zombie twin’s claim is false—has resurfaced. Again, the problem is just the same: if we choose to follow Chalmers here and concede that zombies who are fundamentally mistaken in their self-ascriptive propositional beliefs about the existence of phenomenal consciousness in their own case really are logically conceivable, then we appear to forfeit, once and for all, all hope of motivating any principled distinction between zombie-worlds and the actual world. This, however, would seem to be an absurd conclusion, since it is incompatible with the fact that we already know that phenomenal consciousness really does exist in our world (and, hence, that we are not zombies), purely in virtue of our ability to attend, subjectively, to the qualitative characters of conscious experience in our own case—an ability which, ex hypothesi, simply could not supply equivalent first-person evidential support in the case of our zombie twin, for the trivial reason that, for a zombie, there literally is nothing upon which such an ability might be exercised. After all, for a zombie, “all is dark inside.” And yet, we are told, our zombie twin nevertheless manages to remain “utterly confident” that he/she is phenomenally conscious! As Nigel Thomas puts it, this confidence “is like a stuck fuel gauge that reads FULL whether or not there is any gas in the tank.”32 Fortunately there are, I think, strong additional reasons for thinking that such beings are, in fact, not

coherently conceivable.

2.4 The phenomenology of experiential belief: in defence of Perry’s neo-Humean “epistemology of experience”

John Perry’s critique of the zombie world scenario is straightforward and, to that extent, quite compelling. But before we consider it directly, it will be helpful if we linger, for a moment, on what is arguably the most puzzling aspect of Chalmers’ response to Thomas’s ultimatum. It is important to be clear about what Chalmers’ suggestion actually was. Chalmers’ suggestion was only that direct evidence supplies partial grounds for the justification of our beliefs. It was not the very different claim that it is this same direct evidence of our own phenomenal states that allows our beliefs to coalesce in the first instance. Indeed, given his necessary commitment to the epiphenomenal premise, Chalmers may not say this, since this would be, *inter alia*, to construe conscious experience *itself* as having some necessary (though perhaps not sufficient) causal or explanatory role to play in the very formation of our phenomenal beliefs. That, however, would straightforwardly violate the fundamental claim of the zombie argument, namely, that we can imagine physically identical beings who lack conscious experience entirely, but who are nevertheless functionally equivalent to us. The claim, then, is that the fact of zombie-Adam’s non-consciousness is irrelevant in the explanatory task, undertaken from the third person point of view, of understanding his capacity to form, judge, and to articulate the belief that he is phenomenally conscious. Of course, this also means that the fact of *my* consciousness—and from this, my (true) belief that I am phenomenally conscious—is likewise irrelevant to the functional explanation of the equivalent capacity in me. In light of all that we have seen so far, I think we have good reason to doubt at this point whether Chalmers’ proposed strict

---

33See esp. sections 2.3 of Chapter Two and 3.2 of Chapter Four of Perry, *Knowledge, Possibility, and Consciousness*, called (2.3) “Common Sense about the Mind,” and (3.2) “The Epistemology of Experience.”
separation between first-person knowledge-level justificatory relevance and third-person causal or explanatory irrelevance can actually be sustained. Thus there is, I think, every reason to deny that a zombie, in Chalmers’ sense, really would be functionally equivalent to us, and hence, behaviorally indistinguishable from us from the third-person perspective.

To help bring this out, consider what occurs when one attends, subjectively, to the distinct qualitative aspect(s) of some particular conscious experience. To borrow an example from Perry, take the experience of biting into a warm from the oven, freshly baked chocolate chip cookie for the first time.\(^{34}\) In short, what takes place here (for non-zombies, anyway) is that one is engaged in an intimate, first-personal state of pure qualitative reflection. Reflection upon, that is, the various “raw feels” that constitute the subjective character of the experience itself. Now suppose that our host were to say to us: “So, how was your first chocolate chip cookie?” Our response, perhaps, would likely go something like this (assuming, of course, that we are polite and that we enjoyed our first chocolate chip cookie): “Mmmmm, this is excellent! Why, this may just be the best thing I’ve ever tasted! It’s so warm and gooey; sweet but not too sweet…” and so on. Now, to accept, as a coherent possibility, the fundamental claim advanced by the zombie argument, one must accept that one’s zombie twin, in this same scenario, would exhibit identical enthusiasm, gratitude, and, indeed, pleasure despite the fact that he/she has absolutely no subjective acquaintance—no “direct evidence”—with any of the qualitative characters of the actual cookie-eating experience. Granted, though this scenario may seem strongly counterintuitive, by itself it is not (at least, not yet) obviously fatal to the conceivability of zombies. Perhaps, for instance, we might yet construct a purely functional account of what took place, citing the various physical-electrical-chemical processes that occurred in our zombie-twin’s brain and body, and ultimately (we tell ourselves) succeeding in producing a causally complete description of the admittedly strange behavioral event(s) in question.

However, the plausibility of this picture becomes severely strained once we take a step

\(^{34}\)Ibid, p. 74.
back and consider, on a larger temporal scale, how it could be that our qualia-lacking zombie doppelganger could non-miraculously cultivate a fondness for and a desire to actively seek out, or, at least, to be on the lookout for more, chocolate chip cookie-eating opportunities in the future. Now, for non-zombies who, like me, really do exhibit this sort of long-run behavioral pattern vis-à-vis chocolate chip cookies, arguably the most plausible account that we could give would go as follows. The particular qualitative “impressions” that actual-Adam experienced in eating his first chocolate chip cookie were “coded,” so to speak, into “ideas” which were then stored, indeed flagged, in his memory as being of an especially pleasurable type. Qualia, on this picture, are construed as an integral component of memory-making and, hence, of our capacity to form thoughts about our own subjective experiences. More specifically, they serve as something like the basic “currency” or raw materials of our ability to form certain kinds of beliefs and desires. These, in turn, are what supply us with reasons to engage in the world, as rational actors, in the various ways that we do, as for instance when someone cultivates, and acts from, the qualia-nourished desire to have more chocolate chip cookie-eating experiences.

I believe there is a lot to be said for accepting the broadly Humean story that Perry tells about the epistemology of experience. Moreover, by accepting this account, I believe we find firm ground from which to argue that zombies, in Chalmers’ sense, actually would not claim to be conscious in the full sense—again, for the rather trivial reason that zombies, in virtue of their having no “direct evidence” of conscious experience simply would not be in the right epistemic position (i.e. of first-personal access to/acquaintance with actual qualitative experience) to even form the concept of full consciousness-proper. Perhaps, they would claim to be something like conscious². This, however, is no help to Chalmers, for this would serve merely to underscore the fact that philosophical zombies are neither functionally equivalent to, nor behaviorally indistinguishable from, ordinary humans.

²I am using the terms “impression” and “idea” here in a broadly Humean sense; where “impression” indicates an agent’s direct or “first-order” empirical perception of a given raw experience, and “idea” refers to our “second-order” capacity to form thoughts about our own experiences.
In sum, the strategy I am advocating here can be put as follows: the appropriate response to what Chalmers has called “the paradox of phenomenal judgment” in the case of zombie claims to consciousness is simply to deny that zombies really would claim to be phenomenally conscious, for there are no good reasons for thinking that such beings would (indeed, could), in the first instance, form the second-order belief that they are phenomenally conscious. In a sense, the upshot of this response is to suggest that Chalmers is half right. The crucial point is this. Our direct epistemic relation to our own experiential states is not merely partially justificatory vis-à-vis the reliability of our propositional beliefs about our own consciousness. Rather, it is the brute metaphysical fact that there evidently exists such a relation at all (that is, a relation that holds in the first place) that directly explains how the beliefs themselves are non-miraculously possible in our epistemic economy.

Now, Chalmers might respond at this point by claiming that, at best, my response misses the point, and at worst actually begs the question, since it depends ultimately on what he calls an “inflationary” account of the epistemology of experience. The objection, then, is that I am guilty of ignoring the structure of his argument, a key premise of which, he reminds us, is the stipulation that we should proceed in accordance with a “deflationary” or purely functional notion of belief. The key thought here is that “belief” is something that should be characterized entirely “from the outside,” such that to “believe that-P” (i.e. that one is phenomenally conscious) is simply to exhibit a certain behavioral profile—in this case, the disposition to make certain sorts of claims, inferences, judgements, etc.—which, of course, can be exhaustively analyzed in purely functional terms. According to this notion of belief, there is nothing incoherent in saying that my zombie twin and I are in psychologically identical states—the functional state of believing that-P—because, ex hypothesi, from the third-person point of view my zombie twin and I are externally indistinguishable. By contrast, my account of belief is “inflated” by my suggestion that we need to do more than

---

36This point appears in various places throughout The Conscious Mind; see esp. section 3 of Chapter 1—“Two Concepts of Mind”—called “The double life of mental terms” (see. esp. p.20); see also p. 199 of Chapter 5, “The paradox of Phenomenal Judgement.”
simply attend to what Chalmers alleges we would observe from the outside in this case. Rather, we must first consider what is going on “on the inside,” as this will make a real difference to what we will observe from the outside. This is because my account of what it is to believe \( \text{that-}P \) describes an intrinsically conscious “experiential phenomenon” in its own right, such that for me, or for any other genuine conscious experiencer, to believe, consciously, \( \text{that-}P \) is in fact just to believe \( \text{that one knows-}\text{that-}P \).

Here is one way to frame this distinction. Where the inflationary account is an account of what it is to “believe \( \text{that-}P \),” the deflationary account is actually an account of what it is—or rather, hypothetically what it would have to be—to “believe\(^\ast\) \( \text{that-}P \).” On the former account, which I favour, the fact of conscious experience is itself construed as a necessary though insufficient condition for the possibility of experiential beliefs, and experiential beliefs, in turn, are construed as necessary though perhaps insufficient conditions for self-ascriptive experiential judgments, claims, and so forth. Clearly, the precise nature of the relevant relations here is in need of a fuller explication than I can provide here. Indeed, this is surely one of the fundamental tasks in the construction of a complete theory of consciousness. The point, however, is that such an explanatory account is in principle available in this case. There is, then, the possibility of a naturalistic causal story that we might construct here about how and why conscious experiencers come to believe, and hence to claim, \( \text{that-}P \). Notice that no such account is available in the other case. This is because a zombie’s believing\(^\ast\) \( \text{that-}P \) is, \textit{ex hypothesi}, a phenomenon that cannot be analyzed in terms of, or with reference to, any first-personal acquaintance with qualitative experience internal to the believing\(^\ast\) subject—because, simply, none exists. It would appear, then, that believing\(^\ast\) \( \text{that-}P \) can only be described as a specific behavioral-functional state that zombies, \textit{somehow or other}, just seem to go into from time to time, full stop. Any attempt to understand how and why this occurs would be an explanatory non-starter. All we can say about the phenomenon, really, is that it just sort of happens. In response to this apparent dilemma, my claim is that we should instead simply deny that “believing\(^\ast\) \( \text{that-}P \)” really is
possible.

To clarify, I am not advancing the strong claim that all propositional judgements must ultimately derive from some antecedent causal acquaintance with the object or referent of the judgement in question. I am not, in other words, endorsing a strong or “hard-line” version of the causal theory of reference. Indeed, this position seems clearly untenable. As Chalmers rightly observes, there are some cases of propositional judgements (e.g. religious beliefs, beliefs about UFOs, etc.) where the intentional objects of those judgements are explanatorily irrelevant to the judgements themselves. More to the point, these are judgements that we can arguably explain (away) without any necessary reference to gods or to flying saucers, or what have you. In short, these are quite possibly false beliefs, and hence do not obviously count as instances of knowledge.\(^{37}\) Nor am I suggesting that, unlike zombies who, we are told, mirror this very phenomenon with respect to their false beliefs in their own consciousness, we conscious experiencers are somehow infallible or incorrigible when it comes to our self-ascriptive second-order beliefs about our own phenomenally conscious mental states: surely we are not. But it is one thing to grant that there are bound to be episodes of false or mistaken phenomenal belief in the case of real conscious experiencers. It is something else entirely to claim that the fact of conscious experience itself is wholly irrelevant in explaining how it is that there should even exist the very possibility of (non-miraculous) second-order phenomenal beliefs in the first place. My proposal is that the possibility of self-ascriptive phenomenal beliefs, unlike religious beliefs or beliefs in UFOs, is uniquely causally dependent on the actual existence of phenomenal consciousness in the world. I am not saying that for each and every episode of phenomenal belief by a genuine conscious experiencer, there must be a direct causal link between the phenomenal “object” or, perhaps better, the propositional content of the belief in question and the believing agent. My claim is strictly weaker than this: I say only that there is no prospect for explaining the possibility of deliberate, agentive acts of self-ascriptive phenomenal believing.

\(^{37}\)Ibid. p. 183.
and judging without invoking, at some stage, the causal relevance of an agent’s general cognitive acquaintance with (instances of) the superordinate category of phenomenal experience itself. Granted, the strategy I am recommending may, in the end, still qualify as a weak form of the causal theory of reference, but this alone should hardly be seen as a mark against it, especially since we have rejected the more problematic implications of the theory’s stronger form.

On that note, I should stress at this point that for the purposes of this discussion I am quite happy to accept that zombies might possess, e.g., religious beliefs, beliefs about UFOs, and so forth. Indeed, there is no great mystery as to how this phenomenon should be possible, despite the fact that gods and flying saucers, presumably, would not exist in zombie worlds either. Consider: zombies, as Chalmers describes them, would encounter virtually all of the same gaps, puzzles, and question marks in their explanatory accounts of the observable phenomena in their world as we do in ours—granting, of course, the premise that an absent-qualia physically indiscernible world is itself coherent. Hence, we can explain straightforwardly how these kinds of false beliefs might arise in zombie worlds: namely, both concepts—i.e. that of alien visitors, and that of an omniscient, omnipotent celestial designer (or, perhaps, designers)—would serve an obvious explanatory, or at least descriptive, function in zombie worlds; they are, in sum, theoretical hypotheses. The point I am stressing is that an analogous argument cannot be made in support of the alleged coherence of the proposed presence of the concept of full consciousness proper in zombie worlds. For zombies, there simply is no explanatory or descriptive role whatsoever that this concept could possibly be invoked to play in constructing a theory of consciousness—because, ex hypothesi, there simply is no irreducible, non-functionally analyzable further phenomenon here (no explanandum) that cries out for explanation.

Chalmers observes that, in zombie worlds, “the eliminativists have been right all
along.” On one level, I agree with this. However, the view I am advancing is actually much stronger. My claim is that there really could not be any debate in zombie worlds in the first place. For zombie philosophers, functionalism, of one form or another, was always the only theory of mind that was ever even on the table. Any debate therein could only be about how the fundamental tenets of such a theory, and their relations, should be spelled out. This is because if you are a zombie, then the very notion of *qualia* or of what I have called “full consciousness proper” is, simply, *unimaginable.*

As I see it, the situation here is roughly analogous to an instructive scenario described by Daniel Dennett. Dennett invites us to imagine the case where zombie Martian scientists travel to Earth to try and understand human beings. More specifically, they seek to understand human consciousness. Once they arrive, of course, it is not long before they encounter philosophical qualia-enthusiasts, who naturally try to explain to the Martians “the hard problem of consciousness” as they see it and the “explanatory gap” therein. Unsurprisingly, the zombie Martian scientists are utterly baffled by this. They find all this talk of the *what-it-is-likeness* of conscious experience, the alleged “first-person only” privileged access to “experiential states,” and so on, to be quite unintelligible. Ultimately, Dennett concludes that the sort of third-personal functionalist account that the Martians might go on to construct in this case would be explanatorily exhaustive of human consciousness. In other words, it is not they who have left something out; but rather, we—or, more precisely, the qualia enthusiasts—who have sought, quite illegitimately, to build something in (namely, a by-definition scientifically uninvestigatable “mystery-stuff”—a chimera that does not really exist).

I agree with Chalmers that Dennett’s conclusion here (what Chalmers calls “Type-A materialism”) fails straightforwardly to the extent that it amounts, ultimately, to a denial of the evidence of our own experience. As Chalmers puts it, “this is the sort of thing that

---

can only be done by a philosopher—or by someone else tying themselves in intellectual knots.”

My point is just to say that Chalmers’ imaginary zombie philosophers of mind are in precisely the same situation as Dennett’s imaginary zombie Martian scientists. If we were to imagine somehow visiting a zombie world, or for that matter, zombies coming to visit us (say, for a conference on consciousness!), the qualia-friendly non-zombies would face the very same problem as Dennett’s qualia enthusiasts on Earth. In other words, the conscious experiencers and the zombies would be speaking past each other from the outset, such that any attempt to explain the notion of qualia or of “full consciousness proper” to a zombie would, in effect, be lost in translation virtually before one has even begun to speak.

Nor can we say, for instance, that the fanciful situation described here is exactly equivalent to that of trying to explain, say, the unique tonal quality of a softly played flute to someone who has been deaf from birth. Granted, there are of course some respects in which the two cases surely are similar; and yet, in a crucial sense, the obvious communicative difficulties one might expect to find in the latter case are, I contend, fundamentally less severe than those which would ensue in the imaginary context of one’s trying to explain phenomenal consciousness to a zombie. Consider: it is surely plausible to suppose that an individual who has been deaf from birth would nevertheless still be able to understand the concept that there is something it is like to have an auditory experience of the sort described above (or, indeed, of any other sort). Indeed, there is an obvious and entirely natural explanation as to why this should be so: in short, the reason is because such an individual, presumably, is aware that the vast majority of other people in the world really are in possession of a particular sensory modality that, as it happens, she lacks entirely. But moreover, this individual is, naturally, also already acquainted with the qualitative character of other types of subjective sensory experience; for example, the phenomenology of tasting, touching, seeing, and so on. It is, then, this background phenomenal knowledge that supplies the individual with the crucial framework of reference to help ground and structure our discussion about the

---

40Chalmers, *The Conscious Mind*, p. 188.
phenomenology of sound. Thus it is actually straightforward to see how such an individual would be able to grasp that, had she only been born differently (i.e. with a full set of functioning modalities), she would naturally possess the necessary means with which to undergo phenomenal experiences of an entirely different sort than those which are currently available to her: that is, she would be in a position to have auditory experiences. But again, the mere fact that she is able to self-reflexively comprehend the inherent epistemic limitations of her own situation—to recognize, in other words, that direct acquaintance with a particular sort of qualitative phenomena in the world is simply not available to her—does not, in itself, tell her anything at all about what a flute or a violin sounds like, even if she is able to reason that there is, no doubt, something it is like to experience these, and surely many other sorts of aural phenomena. By contrast, recall that a philosophical zombie is, *ex hypothesi*, a member of a global community, a “zombie world,” that is categorically and uniformly non-phenomenally conscious. A zombie world, in other words, is non-conscious all the way down, such that there is not even the slightest trace or faintest suggestion of qualia anywhere to be found. As such, I contend, quite unlike the case of a deaf or blind human being, a zombie would literally have no idea of what he/she is missing in their engagement with the world.

The core suggestion of this section, then, is not merely the bald assertion that zombies would not claim to be phenomenally conscious, but that the reason behind this is because the hypothetical behavioral-functional state “believing that-P” is itself incoherent, and thus could not occur in a zombie world. More precisely, my fundamental claim is that we cannot really conceive of such a thing as “to believe that-P.” There is only “believing that-P.” The possibility of self-ascriptive belief in one’s own consciousness both depends on and implies some pre-existing cognitive acquaintance with the phenomenon being ascribed: namely, conscious experience itself. Of course, this means that “believing that-P” does not fall neatly into what Chalmers calls the “psychological concept of consciousness,” because experience itself is here construed as explanatorily essential in any
account of how events of this type may even occur. “Believing that-P,” then, is not a purely
“psychological” mental state (in Chalmers’ sense of the term). It is, rather, what we might
call a “psycho-phenomenal” mental state, appropriately analyzable both functionally and
phenomenologically. Indeed, there can be no sharp separation between the two. So un-
derstood, zombies are incapable of believing that-P. The upshot of all this is the following
conclusion. If my imaginary zombie twin and I are first-personally phenomenally distinct
in the way that Chalmers describes (i.e. zombie-Adam lacks conscious experience entirely),
then we are also third-personally behaviorally distinguishable (e.g. I would claim to be
conscious; he would not—and perhaps much else besides). Likewise if my imaginary twin
and I are third-personally indistinguishable and physically indiscernible, then either we are
both zombies, or we are both non-zombies. In either case, the salient point is the same: the
presence or absence of phenomenal consciousness is causally and explanatorily relevant to
our behavioral functioning.

In response to Chalmers' objection, then, I can say only that I do not think that I
have illegitimately assumed an inflationary account of belief, nor that have I begged the
question against his deflationary stipulation. Indeed, in a way my goal throughout this
chapter has been to argue, pace Perry, for the basic soundness of an “inflationary” account
of the epistemology of experience (in Chalmers’ sense of the term) by challenging directly
the sustainability of Chalmers’ proposed bright line separation between the phenomenal and
the psychological concepts of consciousness, and the deflationary, purely functional reading
of beliefs, judgements, claims, and so on, that this implies.

2.5 Conclusion

The overriding conclusion I have been driving at is ultimately to side with Perry and to sug-
gest that, for dualists and physicalists alike, there simply are no good reasons for denying the
“principle of the efficacy of the conscious.” Epiphenomenalism creates far more problems than it solves. Indeed, if nothing else, our examination of the various internal problems with the zombie scenario was intended to demonstrate, above all, that the burden of absurdity that plagues dualist supporters of the zombie argument rests entirely with the demandingness of the conjoined “functionally equivalent/behaviorally indistinguishable” stipulations (and the epiphenomenalism that this presupposes). Stated differently—and speaking, for the moment, from the point of view of someone who, like Chalmers, is already committed to property dualism—there does not appear to be anything prima facie incoherent about the basic notion of a “zombie world”—the notion, that is, of an imaginary world populated by hypothetical beings who are microphysically constituted just as we are, but in which there is no phenomenal consciousness. Again, to borrow a helpful analogy from Perry, so far the zombie scenario I have just described seems every bit as logically conceivable as, for example, the possibility that all the water in the world might (somehow or other) suddenly disappear. Obviously, both scenarios are almost certainly empirical impossibilities. But both seem, nevertheless, to express logical possibilities, in the narrow sense of being internally consistent. Incoherence, it seems, would ensue only in the event that we were to attach two further stipulations to either case. These are: (a) if we were to insist that despite the sudden total disappearance of water from our world, everything else would, nevertheless, remain the same; and (b) if we were to insist that despite the total absence of consciousness (in the full sense) in the zombie world, zombie behavioral functioning would (somehow or other) nevertheless mirror exactly the behavioral functioning in our world. The latter, of

---

41Perry, Knowledge, Possibility, and Consciousness, p. 73.

42As we will examine more closely in subsequent chapters, “antecedent physicalists,” like Perry, would of course deny that this scenario is possible, arguing that conscious experience is itself an “intrinsic” property of the physical, and hence that if the two worlds really were physically identical, then phenomenal consciousness already obtains in both. As Perry puts it, “[T]he subjective characters of our experiences are simply physical properties of physical states.” (Ibid. p.70) The key thought here is that our phenomenal concepts of the qualitative characters of our subjective experiences—e.g. the itchiness of an itch, the painfulness of pain, the redness of red-seeing experiences, etc.—ought to be understood as having their origin in the inherently first-personal phenomenology of what-it-is-like for a conscious subject to be in a particular brain state. I will return to this idea in greater detail shortly.

43Perry, Knowledge, Possibility, and Consciousness, p. 73.
course, is precisely the central thesis of the zombie argument, and that is why it fails.
CHAPTER THREE

Why Mary Couldn’t Know all the Facts

[It should not be supposed that the question of whether Mary learns facts or just apprehends old facts under new guises is simply an issue about how we choose to use the word ‘fact.’ The question is this: is there a way the world is that Mary does not know about until leaving the room...]
—Torin Alter, “A Limited Defense of the Knowledge Argument”

3.1 Taking stock and the move to antecedent physicalism

The primary purpose of this chapter is to examine Frank Jackson’s much discussed “knowledge argument” against physicalism. Before proceeding, it will be helpful to take a moment to consider exactly what we have learned in our investigation of conscious experience so far. To begin, we should isolate the central insight of the preceding chapter. This can be called the “anti-epiphenomenal thesis.” Qualia, we now see, ought not to be conceived as essentially or exclusively qualitative in nature. As Owen Flanagan notes, to insist that they are is
 CHAPTER THREE: WHY MARY COULDN’T KNOW ALL THE FACTS

to endorse the implausible view that qualia “have no other properties than those implicated in their subjectively available aspects” (my emphasis). As the analysis of what Chalmers calls “the paradox of phenomenal judgment” shows—specifically, the pseudo problem of zombie claims to phenomenal self-consciousness—the brute fact of conscious experience itself is inextricably implicated in our understanding of the very possibility of non-miraculous self-ascriptive belief formation in phenomenal self-consciousness. Once this is granted, the conclusion that epiphenomenalism is false is secure: conscious experience is explanatorily essential to a proper understanding of the nature of phenomenal self-knowledge, and hence to the full range of phenomena that makes up our mental lives.

Two things follow from this. We might say they are flip sides of the same coin. The first follows from the brief discussion of Perry, in particular, from the review of his commonsense “neo-Humean” approach of construing qualia as an integral ingredient in memory-making, and hence, as a necessary feature in any plausible theory of the “epistemology of experience.” The salient point here is that any purportedly exhaustive account of the nature of subjective experience framed entirely in descriptive-phenomenological, “behavior-neutral” terms will be seriously incomplete from the start. Though we should not try to eliminate the phenomenological or “private” aspect of experience from our investigation of the subject altogether (as, for example, Dennett seems to recommend), we must nevertheless accept that there is more to the story of subjective experience than a mere inquiry into what a particular quale (e.g. red-seeing, chocolate-tasting, pain-feeling, etc.) feels like or seems like to a conscious subject from the first person perspective. There is also the related question of how the raw impressions themselves relate causally, and hence functionally, to our cognition and behaviour in the various ways that they do.

The second point to note is that any purportedly exhaustive account of the nature of

intentionality that attempts to operate within a purely functionalist, “experience-neutral,” or otherwise deflationary framework inevitably runs aground when confronted with the task of explaining how certain kinds of intentional states (and the behavioral acts that typically accompany them) should even be possible in the first place. As we saw in Chapter Two, paradigmatic of this failure are those intentional states and those behavioral acts having explicitly to do with with an agent’s own phenomenal self-consciousness (i.e. believing, judging, or claiming that-P).

I have also suggested that at this stage in the investigation we should grant, for the sake of argument, the prima facie conceivability of a “zombie world,” even though we have seen strong reasons to doubt the coherence of what we might call a “Chalmers zombie world.” The relevant distinction between these two concepts can be put as follows. The concept of a zombie world refers only to a hypothetical absent-qualia world that is molecule-for-molecule microphysically identical to the real world at “time-\(T_0\).” A Chalmers zombie world, on the other hand, describes a hypothetical absent-qualia world that is microphysically identical to the real world at time-\(T_0\), and remains so at all other times. Perhaps the simplest way to underscore this distinction would be to consider what we would be able to observe if we could somehow hold the two worlds in either hand, thus permitting us to look back and forth between them, to check for fidelity with the actual world. Though both worlds would be objectively indiscernible from the real world at what I have called time-\(T_0\), the zombie world would begin to diverge from the actual world at (roughly speaking) time-\(T_1\), while the Chalmers zombie world, ex hypothesi, would remain perfectly synchronized with the actual world for however long we might look. The relevant distinction, then, is between what we may call a strong absent-qualia hypothesis (i.e. a Chalmers zombie world), and a weak absent-qualia hypothesis (i.e. a zombie world). Now, I have argued that there are good reasons for denying that Chalmers zombie worlds actually are conceivable, and yet the

\[46\]

I borrow this terminology from Perry (2001, Ibid.). Though my use of the term “Chalmers zombie world” directly matches his, I use the term “zombie world” in a slightly different sense than he does. The terms are first introduced on p. 72 of Chapter 4, section 4.1, called “Why Zombies Could Not Be Physically Like Us.”
particular objection developed there does not seem to carry over to the case of a zombie world. Zombie worlds, it seems, are at least antecedently plausible. This is because the anti-epiphenomenal conclusion drawn in Chapter Two is insufficient by itself to overturn property dualism. The analysis that was carried out here sought only to establish the causal and explanatory relevance of phenomenal consciousness. This conclusion, however, remains “topic-neutral” with respect to the further question of whether the metaphysical nature of phenomenal consciousness itself is physical or non-physical. As John Perry notes,

The possibility of zombies, then, seems to be a test for dividing epiphenomenalists from nonepiphenomenalists, not an argument for defending dualism against physicalism. All epiphenomenalists pass the test of finding Chalmers zombies conceivable, either exactly as Chalmers presents them or almost exactly; all nonepiphenomenalists fail it. Both dualists and physicalists pass the test if they are epiphenomenalists and fail it if they are not. . . . [A]t most, then, the zombie argument is an argument for epiphenomenalism.47

Property dualism, then, is actually quite compatible with the impossibility of what we are calling “Chalmers zombie worlds.” Indeed, nonepiphenomenalist dualists and physicalists alike would find common ground in endorsing the critique of this scenario that I presented earlier. Notwithstanding their alignment on this particular point, however, there of course remain deep differences between these two positions—foremost of which for our purposes is that the nonepiphenomenalist dualist might still endorse the weak absent-qualia hypothesis by maintaining that “zombie worlds” are conceptually possible, while the nonepiphenomenalist physicalist, of course, would deny this. This is because the nonepiphenomenal physicalist believes that qualia are themselves to be understood as physical properties of physical states. Hence if the hypothetical world in question really is microphysically identical to the real world at time-\( t_0 \), then phenomenal consciousness already obtains therein, and thus, the hypothetical world is not actually a zombie world. Thus, unlike the nonepiphenomenalist dualist whose only reason for rejecting the possibility of a Chalmers zombie world is because this denies “the principle of the efficacy of the conscious,” the nonepiphenomenalist physicalist additionally maintains that the very

47Perry, *Knowledge, Possibility, and Consciousness*, p. 79.
idea of a philosophical zombie is incoherent, and in this way suggests that even the weak absent-qualia hypothesis can be blocked from ever getting off the ground.

Hence it would appear that we face two possible paths forward: nonepiphenomenal dualism or nonepiphenomenal physicalism. To be sure, our decision about which path we should ultimately take here will depend largely on our examination of the knowledge argument. And yet there are certain other considerations that we would do well to bear in mind even before we proceed. The considerations I am referring to are those concerning the foreseeable consequences of our choosing either path. Of particular interest will be those implications that might be thought to spell trouble for us down the road. With this (conservative) strategy in mind, arguably the most salient implication that we should consider at this point concerns nonepiphenomenal dualism. While it is arguably true that nonepiphenomenal dualism holds a certain intuitive or commonsense appeal at first blush, the cost of actually defending this position is decidedly high. To see that this is so, note that the upshot of endorsing “the principle of the efficacy of the conscious” from a dualist footing is necessarily to commit oneself to having to deny the causal closure of the physical world. Nonepiphenomenal dualism can therefore be seen to introduce a further level of metaphysical complexity—namely, the bifurcation of what there is into two ontologically distinct categories—to the already formidable task of understanding the nature of mental causation. As Perry observes:

The problem with commonsense dualism is not inconsistency but that the arguments for it, however intuitive their force, are simply not compelling in the face of arguments against it. Against it, among other things, is the difficulty of saying much positive and testable about non-physical properties and the wide acceptance of the hypothesis that the physical world is a closed system: that physical events have only physical causes.

Viewed in this light, it becomes clear that the position that we have been calling “nonepiphenomenal dualism” is actually just interactionist dualism by another name. As such, it is straightforwardly vulnerable to the same challenges and criticisms. Foremost

48Ibid. p. 79.
of which, as Perry emphasizes, is simply to observe that before she can make any real progress on the task of providing a satisfying theoretical account of how conscious events make a physical difference in the world, the nonepiphenomenalist dualist must first resolve the arguably deeper metaphysical mystery of how it could even be possible that two fundamentally different kinds of phenomena—the physical and the non-physical—may causally interact with one another in both directions. Granted, the upshot of this exercise is only to try to anticipate some of the major difficulties that would seem to arise downstream from our potentially accepting a dualistic interpretation of the knowledge argument. As such, this obviously does not count as a critique of the argument itself. The point of the exercise, rather, is to suggest that there are compelling, non-arbitrary reasons for entertaining a sort of general sympathy with nonepiphenomenal physicalism going into the argument.

To clarify, I am emphatically not saying that we can somehow establish the decisive failure of the knowledge argument a priori simply by reflecting on the well-known problems of interactionist dualism. Indeed, we may find that the knowledge argument succeeds, in which case the familiar problems of interactionist dualism would be recast into obstacles that would need to be addressed by any theory that attempts to get the metaphysics of mental causation right. My point is only to suggest that from our current fork-in-the-road perspective, we are entirely justified in viewing ourselves as antecedently inclined toward nonepiphenomenal physicalism as our preferred path forward. The position that I am recommending we should adopt at this stage is what Perry calls “antecedent physicalism.” Perry explains,

> Perhaps it would be better to say “prima facie physicalist” rather than “antecedent physicalist.” I do not have in mind a complete dogmatist for whom physicalism is a religious principle...[I] simply mean someone who is committed to physicalism in the sense that she or he sees some compelling reasons for it and will not give it up without seeing some clear reason to do so. The question for such a person is not whether physicalism is the most natural account of the subjective character of experience but whether it offers a possible account.

---

With these thoughts from Perry in mind, we come now to Jackson’s knowledge argument. We should note first, however, that by taking on a sort of general sympathy for antecedent physicalism, we have also thereby assumed a decidedly defensive posture in the overall dualist-physicalist dialectic. This is because from the point of view of antecedent physicalism, the salient question posed by the knowledge argument is whether there are any decisive reasons for thinking that qualia may not be accommodated within a physicalistic framework. The default assumption going in, however, is that this particular demand for accommodation can indeed be met—again, quite straightforwardly, because on this view qualia are themselves to be understood as physical properties of physical states.

### 3.2 Engaging the knowledge argument

In his important paper “Epiphenomenal Qualia,” Frank Jackson develops a compelling argument against physicalism—an argument that he dubs the “knowledge argument.” The goal of the argument is to show that qualia—i.e. the phenomenal or experiential aspects of our mental life—cannot be satisfactorily accounted for by any purely physicalistic theory. The method of the argument is to exploit an intuitive point about the inherent limitations of our epistemic access to certain kinds of information: particularly, information regarding the qualitative or experiential aspects of our own subjective consciousness, as well as the nature of what goes on in the mental lives of others. Ultimately, Jackson’s strategy is to derive from these reflections a strong metaphysical conclusion about the way the world is. Jackson’s central claim is that qualia will always elude capture on any reductive, purely physicalistic theory of the mind. No matter how fine-grained or exhaustive our knowledge of the complex operations of our physical brains and bodies might be, any account of this sort will necessarily leave out crucial facts about the qualitative properties of subjective experience. To help make the point more vivid, Jackson deploys an ingenious thought.

---

experiment and asks us to consider the case of a colour-deprived super scientist named Mary.

Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor. She specializes in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like “red”, “blue”, and so on. She discovers, for example, just which wavelength combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence ‘The sky is blue’... What will happen when Mary is released from her black and white room or is given a colour television monitor? Will she learn anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had all the physical information. Ergo there is more to have than that, and Physicalism is false.\(^{51}\)

The central idea behind the Mary example is to emphasize the fact that there does not appear to be any \textit{a priori} connection, in terms of a logically necessary physical-to-phenomenal entailment relation, between the particular subjective characters of (in this case) colour-quales and the brain states with which they are correlated. If there were such a connection, then physically omniscient Mary would, pre-release, already be in a position to deduce the particular phenomenal feel of, e.g., red-seeing experiences without ever having to exit the black and white room. Mary, of course, is unable to do this; hence, Jackson concludes, physicalism is false. For Mary, it is only by discovering, post-release, what it is like to see red that she is in a position to connect the distinct phenomenal character of \textit{this} subjective experience (i.e. the qualitative “raw feel” of her acquaintance with her first ripe tomato; call this \(R\)) with \textit{that} objective brain state (i.e. the whir of information processing that ensues in our neurophysiological architecture during visual acquaintance with red things; call this \(P\)). Mary, recall, is said to know everything physical that might possibly be known about brain state \(P\) before leaving the Jackson room, and yet before her release she was entirely in the dark concerning the particular qualitative feel of the sensation that people with normal colour vision experience—and which they refer to as “red”—when

\(^{51}\)Ibid. p. 130.
they see, e.g., blood, cherries, ripe tomatoes, and so on. Thus, upon her release, Mary gains *new knowledge* about the mental life of other people that she previously did not possess. In short, Mary learns that events in the brain of physical type $P$ additionally involve *'this'* feeling (i.e. $R$), which Jackson then construes as a further, and, by definition, *non-physical* fact about what goes on in the world. The key point here is that Mary evidently had no epistemic access to a particular sort of information concerning human colour vision before her release, despite the fact that, *ex hypothesi*, she already possessed complete knowledge of all the physical facts associated with the neurophysiology of colour vision. Hence, Jackson concludes, phenomenal properties (e.g. $R$) are ontologically distinct from physical states of the brain (e.g. $P$). Here is the argument in standardized form.\(^{52}\)

Premise 1: Mary has complete *physical knowledge* of all the facts pertaining to human colour vision before her release from the Jackson room.

Premise 2: There is a *kind of knowledge* that Mary does not have before her release — e.g. knowledge of what it is like to see red.

Therefore

Conclusion: There are *non-physical facts* concerning human colour vision (e.g. that $R$ experiences are like *'this'*); hence, not all information is physical information.

There are some fairly strong metaphysical implications that would seem to follow if the knowledge argument actually goes through—that is, in the event that we really are forced to accept the strong dualistic conclusion that the world is incompletely constrained by the totality of all physical information. Perhaps the most obvious implication is that the only way to traverse the "explanatory gap" between the physical and the phenomenal domains, it seems, would be to expand our theoretical framework to include non-physical "bridging

---

\(^{52}\)The rough schema that I provide here borrows heavily from a more expansive prior formulation by Martine Nida-Rümelin. For a more in-depth treatment of the knowledge argument, and a review of the surrounding literature, see her online essay “Qualia: The Knowledge Argument” for the Stanford Encyclopedia of Philosophy: available at: http://plato.stanford.edu/entries/qualia-knowledge/ See also “There’s Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson’s Knowledge Argument,” (The M.I.T. Press, 2004), eds. Peter Ludlow, Yujin Nagasawa and Daniel Stoljar.
principles.” The general strategy, then, would be to introduce new, fundamental laws of nature (i.e. *supervenience laws*) that operate at the “psychophysical nexus” between the physical and the phenomenal, and thus permit us to exert some explanatory grip on the question of why, for example, brain states of physical type $P$ cause us to have $R$ experiences as opposed to, say, $B$ experiences (i.e. phenomenal blueness). Notice, however, that in doing so we would not be positing a form of *logical* supervenience, but rather what Chalmers calls “natural supervenience,” or sometimes “causal supervenience,” which is a weaker relation.

The argument for this proceeds, roughly, as follows. Though Mary, post-release, is now in a position to reflect, for the first time, on the distinct phenomenal character of (certain) colour qualia—and hence to begin to investigate phenomenologically the systematic causal relations that her various textbooks and black and white television lectures had already told her hold between brain states and subjective experience in the world—she is also keenly aware that because the specific phenomenal characters of the quales themselves were only revealed to her *a posteriori* through her direct experience of them, the nature of their relation to the physical, though nomologically firm and consistent, is nevertheless a logically contingent one. Again, if the relation were one of strong logical *necessity*, then Mary could, in principle, already have deduced the raw feel of, e.g. phenomenal redness, blueness, etc., from her complete physical knowledge of the neurophysiology of vision. But of course,

---

53I have no wish to fret over the semantics of whether or not it is proper to describe “black” and “white” as *colors* (perhaps, e.g., it is better to construe these shades, respectively, as different ways in which color might be *absent*). For my purposes, this is a fruitless debate to enter into. Notwithstanding these peripheral concerns, however, notice that it would seem wholly arbitrary to deny that there is nevertheless still *something it is like* for Mary to experience, say, phenomenal blackness and whiteness — and, perhaps, through the exercise of her imagination, various intermediate shades of gray as well — so long as one also maintains that there is something it is like to experience phenomenal redness, blueness, and so on. If this is granted, however, then a dualist proponent of the “explanatory gap” thesis could, arguably, derive the same conclusion from the thought experiment that Jackson does, but in a far more direct manner; namely, without our ever having to imagine Mary being released from the black and white room. Consider: if, *ex hypothesi*, Mary knows all the facts there are to know about the neurophysiology of vision, then it follows that Mary is able to formulate a complete theoretical description of the brain processes that are ultimately responsible for her perceptual capacity to differentiate black things from white things. But more to the point, Mary also knows that there is nothing in this story that implies, much less explains why her visual perception of black things feels the way it does, or why this is phenomenally distinct from the subjective character (the *quale*) that accompanies her visual perception of white things. Indeed, Mary might conceivably also wonder, pace Chalmers, why the functional capacity to make perceptual discriminations of this sort should include any qualitative aspect *at all*. Or so, at least, a dualist might argue.
Mary is unable to do this. Hence, the dualist concludes, any empirical inquiry into the nomically necessary but logically contingent relations that hold between physical states and phenomenal properties in the world will, at best, only reveal something like the “neural correlates of consciousness.” Correlation, however, is not explanation. More precisely, a theory of objective correlations, no matter how fine-grained, will never be fully explanatory (at least, not in the right way) when it comes to answering the further question(s) of how and why subjective experience arises from physical processes in the brain. The explanatory gap remains as wide as ever, and the only way to even begin to bridge this gap, we are told, is by (a) accepting that conscious experience itself is a fundamental property of the world, and (b) by theorizing new psychophysical laws of nature that (somehow or other) govern and facilitate its causal realization. Or so, at least, some dualists have argued.\footnote{See esp. Chalmers, \textit{The Conscious Mind}, pp. 126-128.}

3.3 Physicalist strategies for disarming the knowledge argument

Physicalists have developed a number of different ways of resisting the strong metaphysical conclusion that Jackson seeks to derive from the Mary scenario. The most promising are those that attempt to explain Mary’s new knowledge in a way that is both intuitively plausible and physicalistically acceptable. Less promising, I will suggest, are those that try to deny that Mary acquires knowledge at all. One such attempt is proposed by David Lewis\footnote{D. Lewis, “What Experience Teaches,” in W.G. Lycan (ed.), \textit{Mind and Cognition}, (Oxford: Blackwell, 1990), pp. 499-519.} and by Laurence Nemiro\footnote{L. Nemiro, “Physicalism and the Cognitive Role of Acquaintance,” in W.G. Lycan (ed.), \textit{Mind and Cognition}, (Oxford: Blackwell, 1990), pp. 490-99.}—a strategy that has come to be known as the “Ability Hypothesis.” The central suggestion is that the physicalist should deny that Mary learns anything about the world upon her release, and should instead construe her new knowledge as the acquisition of certain abilities—for example, the ability to recognize, remember, and
(crucially) to imagine, e.g., the particular phenomenal character of red-seeing experiences. The crucial move here is to reject Jackson’s metaphysical claim that Mary comes to know new facts upon her release; hence this position is also sometimes called the “No Propositional Knowledge” view. As Lewis describes it, “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 knowing-that. It’s knowing-how.”\textsuperscript{57} And according to Nemirow, who arguably puts the point in even stronger terms, “knowing what an experience is like is the same as knowing how to imagine having the experience.”\textsuperscript{58} In short, the Lewis-Nemirow conjecture rests on a kind of strong identity claim that asserts the sameness, or at least the epistemic equivalence of (a) imagining what a particular experience is like, and (b) knowing what a particular experience is like.

There are actually a number of different ways that one might go about showing that (a) and (b) are, in fact, importantly non-equivalent. Arguably the most devastating tactic would be to show that simply possessing the conceptual ability to imagine having some particular experience (the colour-quale $R$, for example) is neither necessary nor sufficient, by itself, for knowing what the experience is like. Earl Conee develops this line of argument as follows.\textsuperscript{59} To show that imaginative know-how is not necessary for knowing what, e.g., $R$ is like, Conee asks us to imagine the case of an individual who, for whatever reason, lacks entirely the conceptual ability to imagine any colour experiences. It seems clear, however, that despite this cognitive defect such an individual would nevertheless know what $R$ is like while directly perceiving blood, cherries, ripe tomatoes, and so on. To demonstrate that imaginative know-how is also not sufficient for knowing what it is like, Conee simply inverts this same thought experiment and asks us to instead consider a person (he calls her Martha) who, though perhaps not quite imaginatively omnipotent, is nevertheless something of a savant when it comes to imagining colour experiences. More specifically, Martha is a

\textsuperscript{58}Nemirow, “Physicalism and the Cognitive Role of Acquaintance,” in Lycan (1990) p. 495.
"superlative colour interpolator..." [who is] “highly skilled at visualizing an intermediary shade that she has not experienced between pairs of shades that she has experienced.”

Conee’s Martha example actually has its roots in a much earlier discussion from Hume.

As we touched on briefly in the previous chapter, the central thesis of Hume’s empiricist epistemology is that there are two principal types or “species” of mental perception: thoughts or ideas on the one hand, and impressions on the other. But most important of all, Hume tells us, is to be clear about the dependency relation between the two. Our ability to form a particular thought or idea about something is to be understood as parasitic on our having had some prior impression(s)—via our direct empirical acquaintance with (aspects of) the world—of the particular entity or phenomena in question; i.e. that which the thought is about, or is an idea of. According to this view, a man who is blind from birth, for example, would be deemed entirely unable to form colour thoughts because, lacking vision, he lacks the requisite sensory modality with which to acquire the antecedently necessary impressions of colour in his perception of the world. There is, however, a certain gap in Hume’s overall account, which Hume recognizes:

Suppose...a person to have enjoyed his sight for thirty years, and to have become perfectly acquainted with colours of all kinds except one particular shade of blue... which it has never been his fortune to meet with. Let all the different shades of that colour, except that single one, be placed before him, descending gradually from the deepest to the lightest; it is plain that he will perceive a blank, where the shade is wanting, and will be sensible that there is a greater distance in that place between the contiguous colours than in any other. Now I ask, whether it be possible for him, from his own imagination, to supply this deficiency, and to raise up to himself the idea of that particular shade, though it had never been conveyed to him by his senses? I believe there are few but will be of opinion that he can; and this may serve as a proof that the simple ideas are not always, in every instance, derived from correspondent impressions; though, this instance is so singular, that it is scarcely worth our observing, and does not merit that for it alone we should alter our general maxim. (Hume, 1777, section 2, par. 16)

Setting aside the last line from this passage, which is obviously controversial, it is clear that Conee’s example of Martha the “superlative colour interpolator” trades on this very

---

60 Ibid. p. 138.
point from Hume. But what is most important to our current purposes is to note that Conee could well concede to the Lewis-Nemirow conjecture that, for instance, Martha’s capacity to visualize Hume’s missing shade of blue—call this $Bm$—is potentially sufficient for her knowing what $Bm$ is like, and yet still deny that her mere possession of this imaginative ability is itself identical or equivalent to knowing what it is like. As Conee explains, insofar as Martha does not exercise her ability to visualize $Bm$ (for whatever reason), she simply does not know what it is like to experience $Bm$—notwithstanding the fact that knowledge of the latter sort is, in principle, epistemically accessible to her by an alternative means than that of the standard eye-to-brain direct perception of an object of the appropriate shade. The upshot of our review of Conee’s critique, then, is simply to observe that though there would seem to be good reasons for granting Hume’s conjecture that someone like Martha might plausibly come to know what $Bm$ is like by non-ocular means, the mere possession of the ability to perform this feat does not, in itself, amount to knowing what $Bm$ is like.

So much for the Ability Hypothesis. It is worth noting that Conee also provides an interesting positive proposal for how he thinks a physicalist should deal with the Mary scenario—a proposal that is sometimes called the “Acquaintance Hypothesis.” Conee’s suggestion is that the physicalist should construe Mary’s new knowledge of $R$ as a non-propositional addition to, or, perhaps better, an extension of her prior knowledge (which is propositional) that “seeing red things” causes persons with normal colour vision to have experiences with a particular phenomenal character. According to this view, because she possesses a complete theoretical understanding of the neurophysiology of vision before her release, Mary already knows under what conditions phenomenal redness occurs (namely, during red-channel activation in our brain); she just doesn’t know what phenomenal redness feels like. Thus we may plausibly say that, pre-release, Mary possesses something like a theoretical concept of phenomenal redness. Call this $QR$. Now, because Mary’s red-channel has never actually been turned on—whether by her directly perceiving something red or, perhaps, by her somehow manifesting and exercising Martha-like imaginative abilities—the
status of $QR$ in Mary’s cognitive economy is rather like that of an empty slot in a file-folder or computer cache. There is, then, a spot and a label for the property “phenomenal redness” in Mary’s pre-release understanding of colour vision. It’s just that this is destined to remain empty until such time as she is actually released from the Jackson room (or is given a colour television monitor). $QR$, then, can be understood as a sort of “epistemic placeholder” that lacks the appropriate corresponding phenomenological component or recognitional core. Of course, once Mary is released and sees her first ripe tomato, $QR$, in effect, becomes $R$.

According to Conee, however, there is nothing metaphysically problematic, much less anti-physicalistic, about Mary’s epistemic progress here. This is because Mary already knows that $R$ perception is caused by red-channel activation in the brain, which can of course be exhaustively described in physicalistic terms. Conee’s point is to emphasize that because we know that visual experience is itself a physical process involving complex relations between the eye and the brain, colour qualia, which we take to be properties of visual experience, should themselves be understood as physical in kind. From here the task becomes that of explaining how, exactly, Mary is able to move, epistemically, from theoretical concept $QR$ to phenomenal concept $R$. But again, this need not be seen as especially troubling for the physicalist. The appropriate response, Conee suggests, is just to say that Mary has come to know something by acquaintance. Knowing something by acquaintance (in this case, phenomenal redness) “requires the person to be familiar with the known entity in the most direct way that it is possible for a person to be aware of that thing.”

In some ways, this remark intersects with a related point from William Lycan, who suggests that a concept has, on the one hand, an “extension” (by which he means, “the set of items to which the concept applies,” see p. 118—in Mary’s case, this “set” consists in all the relevant low-level physical facts associated with red channel activation); and, on the

---

other hand, a “functional profile,” by which he appears to mean (something like) the possible intentional modes in which the concept might figure in its owner’s psychology. Lycan suggests that the two ought to be understood as strictly independent of each other, and that this supports the conclusion that, as he puts it, “knowledge is finicky and hyperintensional” and that “(not surprisingly) knowledge involves the mode under which the knower represents the fact known.”

Thus, to “know” that red perception involves the experience of phenomenal redness under the purely extensional mode of the concept $QR$ is importantly epistemically distinct from one’s employing $R$ (which, in a sense, is the explicitly phenomenological or first-personal “mode” of $QR$) in speaking, or in thinking, about the same item of information. There is, however, no metaphysical or ontological import to any of this, since $QR$ and $R$ are, simply, different ways or “modes of presentation” under which one may come to know, and think about, a single fact. In sum, Conee’s claim is that Mary’s coming to know phenomenal redness by direct acquaintance need not be taken to imply that she has acquired any new propositional knowledge, because all the facts about red-seeing were previously known to her under her complete understanding of the neurophysiology of visual experience—which, we are to recall, included the theoretical concept $QR$.

Conee is in fact only half right on this point. To see where the acquaintance hypothesis fails, consider that on this view knowledge by acquaintance is construed as a third category of knowledge unto itself, irreducibly distinct from both ability knowledge and propositional knowledge. Thus, according to this view, $QR$ and $R$ may convey different semantic content, even though both concepts actually designate the same thing—namely, phenomenal redness. The relevant difference is that where $QR$ describes the phenomenal character of red-seeing indirectly, from the third person perspective, $R$ involves what Terrence Horgan calls “the first person ostensive perspective on that property,” and hence picks out phenomenal redness directly.

In short, $QR$ and $R$ “co-refer” to phenomenal redness. More precisely, Mary’s

---

64 Ibid. p. 113.
“$R$ knowledge” does not correspond to any new facts about the particular property in question—all of these are already subsumed under $QR$.

It seems clear that something fundamental has been left out or swept under the rug with this account. The problem lies with an incomplete picture of all that takes place during Mary’s epistemic progress from $QR$ to $R$. To help bring this out, consider that $QR$ is what we might call a third personal theoretical concept of an inherently first personal property. As such, it lacks a phenomenological core (as Owen Flanagan puts its, $QR$ lacks “the phenomenal component of the concept of red.”66). What this means, then, is that $QR$ does not truly refer to what the experience of seeing red is actually like for people with normal colour vision. It merely signifies that seeing red is like something for normal perceivers. Ultimately, of course, Mary does move from $QR$ to $R$. Thus Mary comes to know what it is like, in the phenomenological sense, to see red. From here, of course, the salient question becomes whether Mary thereby acquires information that corresponds to any new facts. We have also seen that both the Ability Hypothesis and the Acquaintance Hypothesis deny that Mary learns anything about the world. Intuitively, however, this assertion seems distinctly unsatisfying. Let us now consider the alternative.

### 3.4 On knowing what it’s like: acquaintance plus recognition

To see that Mary’s new knowledge is genuinely propositional, consider Mary’s own pre-release perspective on the phenomenology of colour. More precisely, consider her perspective on the specific qualitative characters of human colour perception that actually obtain in the outside world. Considered from this point of view, it seems clear that Mary, pre-release, simply has no idea what the phenomenal referents of terms like “red,” “blue,” “yellow,” and so on actually are in the world. This is because, pre-release, these terms—and the scientific

---

descriptions of the various brain states with which they are systematically correlated—simply do not (yet) contain a crucial item of further information—namely, phenomenal information. Indeed, as far as Mary knows, the actual phenomenal referents of these terms could be anything—or, more precisely, they could be anywhere from within the fixed set of possible spectral discriminations that human vision is, in fact, able to perceive in light—but of course Mary already knows this. Again, as far as Mary knows pre-release, the actual phenomenal referent of “red,” or what we have termed the phenomenological component or “core” of QR, might be like ‘this’ or it might be like ‘that.’ Thus, Mary’s discovery of what red-seeing is like—or, alternately phrased, her discovery that red-seeing is like ‘this’ and not like ‘that’—is, straightforwardly, a discovery about the world. Moreover, given that she is now able to reflect on the distinct, sui generis “this-quality” of phenomenal redness, Mary is also now in a position to form genuinely new thoughts and beliefs about red perception, thoughts and beliefs which were simply not available to her under her previous explicitly physicalistic conceptualization. The thought here, roughly, is that Mary’s progress from what we are calling QR to R consists in her narrowing down the space of epistemic possibilities—possibilities which, though not exactly boundless (for the reasons mentioned above), were nevertheless indeterminate before her release.

This is a crucial point, though it is also one that is easy to misunderstand or to misinterpret. Let me be clear. I am not saying that the actual phenomenal referent of the theoretical concept QR (namely, what red-seeing is, in fact, like for normal perceivers) should in any way be thought of as metaphysically indeterminate. Indeed, if we were to say this, then we could no longer maintain that qualia are, at bottom, physical properties of physical states. My point is simply to stress that despite their being an objective, metaphysical fact of the matter to this question—roughly speaking, that, given the way the world is, red channel activation or what I have called “brain state P” causes us to undergo subjective experiences of phenomenal type R—this nevertheless is not a fact that is knowable a priori.
To help bring this out, consider the following variation on Jackson’s original thought-experiment (here I follow Martine Nida-Rümelin, who first developed this idea in her 1995 paper “What Mary Couldn’t Know: Belief about Phenomenal States”). For the sake of argument, suppose that Mary was not released from the black and white room directly, but was instead taken to an intermediary room or “staging area” first. Call this the Nida-Rümelin room. The Nida-Rümelin room, let us suppose, is windowless and contains no furniture or objects of any kind (save for whatever source of colour-neutral illumination was provided in the black and white room). Indeed, the only feature of this room that distinguishes it from the black and white room (apart from its being completely empty) is that two of its four walls are painted red, and the other two are painted blue. Now suppose that Mary is told this before she goes inside, and has no reason to doubt the source. What she is not told, of course, is which of the four walls are red and which are blue—though, as it happens, walls 1 and 2 (suppose they are numerically labeled) are both red; walls 3 and 4, then, are both blue. Once she is inside, Mary might go on to form two formal concepts to assist in her thinking; concepts whose function is to succinctly describe and thus differentiate the two distinct colour impressions with which she is directly acquainted: walls 1 and 2, let us say, she terms “Q-this”; walls 3 and 4 she terms “Q-that.” Now consider what else Mary knows. Mary knows that one of the two shades before her is “red,” or what I am calling “the phenomenal referent of $QR$,” and she also knows that one of the two shades before her is “blue,” or what we might call “the phenomenal referent of $QB$,” which is epistemically equivalent to $QR$. She just doesn’t know which one is which.

Thus, from Mary’s epistemic perspective within the Nida-Rümelin room, both of the following propositions express coherently conceivable possibilities about the world: namely, that “Q-this is $QR$” or that “Q-that is $QR$”—and of course the same can be said with respect to the phenomenal referent of $QB$. The point, of course, is that Mary has no means

---

67My treatment here is essentially an extension of a similar scenario discussed by John Perry (Idem. see esp. pp. 111-13, and 148-50.), though Perry, in turn, adapts his discussion here from a scenario that was originally formulated by Nida-Rümelin in “What Mary Couldn’t Know: Belief about Phenomenal States,” in Metzinger, 1995: pp. 219–241.
whatsoever from within the Nida-Rümelin room of ascertaining the truth-value of these propositions. Notice also that despite the fact that she is directly acquainted with phenomenal redness while attending to the raw qualitative feel of \( Q\)-this, Mary still does not truly know what it is like to experience phenomenal redness because she does not know that \( Q\)-this is \( QR\). In short, Mary doesn’t know which of the two identity statements in question corresponds to the way the world actually is. Granted, Mary does know that phenomenal redness is either like \( Q\)-this or \( Q\)-that—ditto phenomenal blueness. Formally, we can state Mary’s epistemic situation here in the form of a disjunction between the following two conditionals.

Proposition 1: if (\( Q\)-this is \( QR\)) then (\( Q\)-that is \( QB\)).

Or

Proposition 2: if (\( Q\)-this is \( QB\)) then (\( Q\)-that is \( QR\)).

Though both are coherently conceivable for Mary from within the Nida-Rümelin room, only one describes the state of affairs that actually obtains in the world. Mary knows this. From Mary’s perspective, of course, neither is any more or less likely than the other to turn out to be true. Indeed, the possibility that, for example, gazing out over the ocean is like \( Q\)-this is, for Mary, entertained as being exactly equally plausible as that it is like \( Q\)-that. More formally, we may say that each thought—or, perhaps better, each “entertaining”—has equal \textit{a priori} justification from within the Nida-Rümelin room. Thus, as for which proposition is true and which is false, Mary is necessarily agnostic—at least until she is released into the outside world.

The crucial point to note here is that direct acquaintance, though necessary for knowing what it’s like, is clearly not sufficient for knowing what it’s like. At the risk of abusing the language even further, absent what we might call a “recognitional core” and direct acquaintance only tells one what \textit{something} is like. Thus, to know what \textit{it} is like (in this case, phenomenal redness or blueness) requires the conjunction of direct acquaintance with
CHAPTER THREE: WHY MARY COULDN’T KNOW ALL THE FACTS

recognition—which is of course precisely what occurs in Mary’s epistemic economy when she is released from the Nida-Rümelin room and encounters, for the first time, ripe tomatoes and oceanic vistas. In sum, Mary learns that $Q$-this is $QR$ and that $Q$-that is $QB$, which in turn gives her the pure phenomenal concepts $R$ and $B$ respectively. As John Perry describes the situation, “[T]he space of what is conceivable for Mary will have changed as a result of her new knowledge. What is conceivable for Mary now coincides with what is possible.” [For example, the hypothesis that $Q$-that is $QR$ is no longer conceivable.] And, Perry continues, “Notice that there is no way Mary could have taken this step a priori.”

With the acquaintance-plus-recognition framework now in place, in order to forestall a natural objection to my account from certain physicalists, there remains a crucial further point of terminological clarification that needs to be stressed. The objection I am referring to comes from the “no new propositional knowledge” physicalist who might well cite at this point what she perceives as an ambiguous or under-specified usage on my part of the term “fact”—a usage which, she might argue, not only underlies my claim that Mary’s new knowledge is genuinely propositional, but that ultimately undermines it. To the contrary, my use of this term is entirely straightforward. As Torin Alter observes, “[I]t should not be supposed that the question of whether Mary learns facts or just apprehends old facts under new guises is simply an issue about how we choose to use the word ‘fact.’ The question is this: is there a way the world is that Mary does not know about until leaving the [black and white] room...” My central claim throughout this section is that the physicalist should respond in the affirmative to this question and that in doing so she thereby grants both of the following: (a) she accepts that Mary learns new facts upon her release, and (b) she accepts that not all facts about conscious experience are discursively learnable. Conversely—and this is surely the most common physicalist move in the debate—to respond in the negative to Alter’s question is to insist not only that, pre-release, Mary

---

68 Perry, Knowledge, Possibility, and Consciousness, p. 149.
possessed all the facts about colour experiences, it is also to insist, I think quite implausibly, that it is a necessary condition of something’s even counting as a ‘physical fact’ that the information it contains must be discursively learnable. In sum, my use of the term “fact” in this debate is meant to refer only to the possibility of someone’s possessing, or coming to possess, information in virtue of which that person might properly be said to know something that corresponds to a truth, or to truths, about the way the world is. So understood, discursive learnability need not and should not be taken as an analytic test-standard for whether epistemic content is factual or not. Indeed, as the Mary example shows, there are discursively unlearnable facts about the world. The salient point here can be put linguistically, in terms of which statements Mary can and cannot understand. First, recall that Mary is able to effectively employ the third-personal theoretical concept $QR$ in both her pre-release thinking about colour vision, as well as in technical conversations with colour-sighted brain scientists from the outside world. The content of $QR$, recall, is something like this: “phenomenal redness is the property that people with normal colour vision experience during red channel activation in the brain.” It is, in other words, what I have called an “epistemic placeholder.” To be sure, so far as it goes, $QR$ clearly does express facts about red-seeing. However, it is also clear that $QR$ does not exhaust all the potentially knowable facts there are to learn about red-seeing, since it fails to convey what red-seeing is actually like. Nevertheless, at one level perhaps we might still be inclined to say that Mary ‘understands’ phenomenal redness. But we need to be careful here in how we are using the term ‘understand.’ Consider: it seems clear that the only sense in which Mary’s ‘understanding’ of phenomenal redness might plausibly be said to match that of the colour-sighted brain scientists with whom she sometimes corresponds is that she too is able to correctly refer, by name, to the appropriate neurophysiological concept in theoretical statements regarding red channel activation. This sort of mastery of correct theoretical concept use, however, is clearly not sufficient for complete understanding, since

---

70 Ibid. See esp. pp. 50-51.
71 Ibid. p. 48.
it need not imply that the concept user has any grasp of the full or specific nature of the actual property or properties that the concept in question is employed to denote. After all, I might conceivably use the terms “nirvana” or “enlightenment” correctly, in a purely contextual sense, when discussing the tenets of Buddhism with the Dalai Lama, and yet still not truly understand the full extent of the complex ideas, attitudes, practices, and so on, that I nevertheless successfully manage to refer to. Indeed, the point seems only more vivid in Mary’s case with respect to phenomenal redness. As Alter puts it, “She does not comprehend what it is for an experience to have that property, and so she should not be described as knowing the fact that red sensations are phenomenally red” (my emphasis). 72

This last line, of course, is simply a different way of stating the fact that QR is Q-this, which, inter alia, is precisely what Mary learns once she is released from the Nida-Rümelin room.

3.5 Flanagan on the distinction between linguistic and

metaphysical physicalism 73

The upshot of the previous section was to emphasize that Mary would credibly claim to have learned something new about colour perception upon her release from the Jackson room—and, hence, about the world—and that this is something that physicalists need to face up to squarely. There is, of course, a strong tendency here to deny that this sort of accommodation can plausibly be made from a physicalist footing; to insist, in other words, that one really cannot have it both ways. The allegation here is that by granting that Mary discovers something about the world, the physicalist necessarily gives up their central thesis (even if they do not immediately realize this). To see that this is not so, that allegations of this sort actually rely, crucially, on a caricatured account of physicalism and thus fail to hit

72 Ibid. p. 48.
73 Flanagan, Consciousness Reconsidered, see esp. Chapter Five, pp. 87-102.
their intended mark, consider the following remarks from Owen Flanagan—remarks which, I will suggest, are indispensable in our getting clear about what, precisely, the doctrine of physicalism actually is, and thus just what it is that we are “antecedently” committed to defending.

One way to avoid the seduction here is to distinguish between metaphysical physicalism and linguistic physicalism. Metaphysical physicalism simply asserts that what there is, and all there is, is physical stuff and its relations. Linguistic physicalism is the thesis that everything physical can be expressed or captured in the languages of the basic sciences: “completed physics, chemistry, and neurophysiology.” Linguistic physicalism is stronger than metaphysical physicalism and less plausible (Fodor, 1981). Jackson gives Mary all knowledge expressible in the basic sciences, and he stresses that for physicalism to be true all facts must be expressed or expressible in “explicitly physical language.” This is linguistic physicalism. It can be false without metaphysical physicalism being false. This is easy to see. 74

Considered in this light, we are now able to see where, on the one hand, the knowledge argument first goes wrong, but also—and this is surely more interesting—where and why the standard objections to the argument fail. Both the Ability Hypothesis and the Acquaintance Hypothesis, we now see, object to what they regard as Jackson’s illegitimate inference from premise 2—that there is a kind of knowledge that Mary does not possess—to the strong metaphysical conclusion that the world is, in part, made up of non-physical facts (i.e. that “not all information is physical information”). It turns out, however, that the real problem all along lies not with any alleged invalid inference on Jackson’s part, but rather with the charity of certain physicalists who are inclined, far too quickly, to grant Jackson’s first premise—that Mary has complete physical knowledge. Mary does not have complete physical knowledge before her release from the black and white room, because she lacks crucial phenomenal information about the subjective character of colour experiences. Phenomenal information is still, however, coherently a form of physical information. It is, as Terrence Horgan puts it, a form of “ontologically physical information,” as opposed to “explicitly physical information.” 75 There is, then, no incompatibility or inconsistency,

74 Ibid. p. 98.
75 The relevant passage is on p. 150 of Horgan’s paper “Jackson on Phenomenal Information and Qualia,” Philosophical Quarterly 32: 147–152. Here, Horgan writes, “In order to develop this point, we need to
much less formal contradiction in the physicalist asserting that despite Mary’s possession, pre-release, of complete explicitly physical information, she nevertheless comes to learn something new about the world upon her release. Indeed, as Flanagan further observes, I think quite insightfully, physicalists might also point out that, pre-release, Mary is only too aware that she would be in a position to discover new facts about the world, if only she were to be released. But moreover, given her complete theoretical understanding of how the neurophysiology of vision works, Mary is also entirely well-equipped from within the Jackson room to explain precisely why this should be so: namely, by citing what we might now regard as the rather banal fact that “perceptual capture is done in the first person.”

Thus, it would seem, physicalism can indeed have its cake and eat it too.

characterize the two relevant senses of ‘physical information.” Let S be a sentence that expresses information about processes of a certain specific kind, such as human perceptual processes. We shall say that S expresses explicitly physical information just in case S belongs to, or follows from, a theoretically adequate physical account of those processes. And we shall say that S expresses ontologically physical information just in case (i) all the entities referred to or quantified over in S are physical entities, and (ii) all the properties and relations expressed by the predicates in S are physical properties and relations. Thus, explicitly physical information is expressed in overtly physicalistic language, whereas ontologically physical information can be expressed by other sorts of language — for instance, mentalistic language.”

76Flanagan, Consciousness Reconsidered, p. 100.

77My phrasing here is a direct reference to the following line from p. 168 of Chalmers’ The Conscious Mind: “The moral is that those who want to come to grips with the phenomenon must embrace a form of dualism. One might say: You can’t have your materialist cake and eat your consciousness too.”
Chapter Four

Conclusion

4.1 In defence of antecedent physicalism

The chief task of this section will be to draw together the conclusions from our investigation so far in an effort to state the most plausible stance that one might take in the contemporary dualist-physicalist dialectic surrounding the nature of phenomenal consciousness. To be sure, both analyses of the two major anti-physicalist arguments provided here have been motivated by a general skepticism about the dualist’s project. In this sense, the overall tenor of the discussion throughout might fairly be described as one of anti-dualism. Even so, there remains an important sense in which the fundamental aim of this project is actually better described as a defence of principled agnosticism. Granted, I have indeed suggested, pace Perry, that given the current state of the debate the most plausible position to affirm is that of “antecedent physicalism.” And yet, as we have also seen, this posture is importantly distinct from that of endorsing or, perhaps better, crediting physicalism as the canonical truth of the matter. This is because the antecedent physicalist does not so much endorse physicalism as she rejects dualism—or rather she rejects the arguments currently on offer.
in support of dualism, as failing to make a decisive case against physicalism, properly understood. This leaves the \textit{prima facie} plausibility of physicalism intact and thus, I contend, antecedent physicalism remains the appropriate default position in the debate.

In a sense, one might fairly describe antecedent physicalism as a kind of “negative physicalism” within the internal dialectic of physicalism itself. Conversely, the views clustered at the other end of the spectrum are what might be called the “positive physicalisms.” These are the views whose chief proponents tend to argue that philosophy should take a backseat in the task of constructing a proper scientific theory of consciousness.\footnote{Daniel Dennett and Patricia Churchland are paradigmatic examples of what I am calling “positive physicalists.” See esp. Dennett’s paper, “Facing Backwards on the Problem of Consciousness” (pp. 33-37); see also Churchland’s “The Hornswoggle Problem” (pp. 37-45). Both are contained in the collection \textit{Explaining Consciousness: The Hard Problem}, ed. Jonathan Shear, (The M.I.T. Press, 1999).} In contrast to these more strident physicalisms, the chief concern of the antecedent physicalist is to demonstrate systematically why the standard anti-physicalist arguments do not go through. This is the sense in which I have described the posture of antecedent physicalism as largely defensive in nature. With that in mind, I would stress that the primary goal of my project throughout has not been to \textit{prove} dualism false, any more than it has been to \textit{prove} physicalism true. The goal, rather, has been to provide principled reasons for thinking that, other things being equal, physicalism is the view that currently enjoys the most evidence, and has the most philosophical justification in its support. Thus, I contend, it is here that we should place any long-term bets concerning the metaphysics of mind.

To show that the deck really is stacked against dualism, we considered the two major arguments against physicalism, the zombie argument and the knowledge argument, and found both to be unpersuasive. To begin, let us briefly recap our review of Chapter Three. Here we saw that the central failing of the knowledge argument is its unjust attempt to saddle the physicalist with the hopeless task of having to defend the following thesis, which Flanagan calls “linguistic physicalism.” Namely, that it is a necessary condition of physicalism’s being true not only that there actually exist an objective implication from “low-level” micro-facts
about the brain and the body to “high-level” macro-facts about subjective experience, but
that the latter must also be fully conceptually derivable a priori, in all their qualitative
specificity, purely in virtue of one’s possessing a systematic theoretical understanding of
the former. The main problem with this strategy, as we have just seen, is that linguistic
physicalism need not be true for metaphysical physicalism to be true. Indeed, metaphysical
physicalism, properly understood, is simply the monistic view that “what there is, and all
there is, is physical stuff and its relations.”79 Thus full describability in explicitly physi-
calistic language—or EPL for short—just is not an analytic test-standard for determining
whether or not something is ontologically physical in kind. Similarly, as was shown from
our brief discussion of Alter, neither should the physicalist concede to dualists (or, for that
matter, to other physicalists) that discursive learnability is a plausible test-standard for
judging whether something should count as a ‘physical fact’ or not. If metaphysical physi-
calism is true, then all facts are ‘physical facts,’ up to and including those that Mary only
learns a posteriori, once she is released from the Jackson room. As was shown, by far the
more plausible reading of factual knowledge is simply to say “information in virtue of which
a person might properly be said to know something about the way the world is.” The
chief advantage of this account over the conventional physicalist view, of course, is that a
posteriori direct experience is openly acknowledged as a necessary condition of an agent’s
possible epistemic access to otherwise inaccessible “ontologically physical information.”

We also noted from our discussion of Conee that in light of all we know concerning
the myriad physical processes that together constitute visual perception itself—i.e. “all
the chemical processes in the rods and cones (the place where the differences in colour
experiences that we know of have their origin) and beyond, including events in the visual
cortex and anywhere else relevant to vision and the experience of it”80—it would seem at
best mysterious, and at worst perverse to insist that colour qualia, which we take to be
properties of visual perception, must nevertheless be understood as ontologically distinct

79Flanagan, Consciousness Reconsidered, p. 98.
80Perry, Knowledge, Possibility, and Consciousness, p. 91.
in kind from their physical realization. Once it is made clear that the truth of physicalism does not conceptually require either \textit{EPL} or \textit{DL} (i.e. the “discursive learnability thesis”) this suggestion loses much of its intuitive force.

Finally, the fact that Mary’s new knowledge that \(QR\) is \(R\) is an \textit{a posteriori} discovery that, accordingly, can only be made once she is actually released into the world does not, by itself, license the strong dualistic conjecture that \(R\) experiences must in some sense be taken to correspond to nonphysical facts about the way the world is. Indeed, as we have just seen from our discussion of Flanagan, Horgan, and Alter, there is another strategy here as to how Mary’s discovery might plausibly be construed. To borrow a phrase from Chalmers, there is an interpretation that is far less “ontologically expensive” than that of dualism.\(^{81}\) This strategy, in short, is just to say that though \(R\) sensation is indeed metaphysically necessitated by brain state \(P\)—meaning, roughly, that the internal state “red channels on” \textit{ensures} that one is experiencing phenomenal redness and not, say, greenness—by its very nature, the qualitative essence of \(R\), being a phenomenological “felt-state,” just is not the sort of thing that is ever fully \textit{revealed} until the \(R\) perceiving subject is actually \textit{in} brain state \(P\). Once again, the physicalist explanation for this is straightforward: “perceptual capture is done in the first person;” full stop. The core suggestion, then, is that the entailment relation between complex microphysical states of affairs in our brains and the phenomenal experiences that they instantiate in us—and, crucially, the precise nature and

\(^{81}\)See esp. pp. 388-89, of “Moving Forward on the Problem of Consciousness,” in \textit{Explaining Consciousness: The Hard Problem}, ed. Jonathan Shear, (The M.I.T. Press, 1999). Chalmers uses this language to criticize what he sees as the failure of the identity thesis to serve as a kind of “escape clause” for what he terms type-B materialism. Chalmers suggests that what is really going on here is that one is trying to get a primitive identity statement to “do all the explanatory work of a fundamental law at none of the ontological cost.” The identity theorist, in other words, is trying to get something on the cheap. Chalmers suggests that “we should be suspicious of such free lunches.” My central claim throughout this, and the latter sections of the previous chapter, is that once physicalism has been separated from the misguided theoretical trappings of its conventional \textit{linguistic} incarnation (i.e. once the \textit{EPL} and \textit{DL} theses have been removed), this sort of objection loses much of its critical bite. If metaphysical physicalism is true, then an identity thesis is an entirely plausible conjecture as to how we might explain the fixed correlations that exist between explicitly physicalistically describable states of affairs in the brain (e.g. “red channels on”) and the \textit{non}-explicitly physicalistically describable, \textit{a priori} underviable, subjective experiences that they instantiate in us (e.g. the \textit{phenomenal redness} of red-seeing). In sum, the fact that the specific subjective characters of experiential states cannot be deduced \textit{a priori} purely from one’s having complete “explicitly physical information” is certainly a problem for \textit{linguistic} physicalism, but it is not a problem for \textit{metaphysical} physicalism.
implementational details of the actual realization mechanism(s)—just is not epistemically transparent to us, at least not yet.

Thus, in a sense, the hard problem of consciousness persists; the explanatory gap remains. The point that I am pressing, however, is that dualists are not justified in drawing the strong metaphysical conclusion that we are actually facing an *ontological* gap here, as opposed to a merely epistemic one. It is simply far too early to conclude that the hard problem of consciousness *may not* be solved by further empirical investigation. Doubtless the dualist would reply to this remark by noting that the “wait and see” strategy that I seem to be recommending in the very least acknowledges the possible existence of a certain “psychophysical nexus” between brain states and conscious experience. For the sake of argument, the antecedent physicalist may go along with this reading. The problem for the dualist critic, however, is that by itself this is not a compelling objection to much of anything. In response, the antecedent physicalist need only stress that dualists are premature in concluding that any plausible account of the realization process that occurs at this “nexus” (meaning, this *epistemic* blind spot) *must* be construed in quasi-alchemical terms as a “conversion-into” or “creation-from” that which is physical to that which is not. To be sure, this is one of the quintessential dualist non sequiturs in the entire debate. Indeed, if the metaphysical physicalist is right and qualia are, at bottom, physical properties of physical states, then the precise realization mechanism(s) in question could, in principle, eventually be recognized and understood in much the same way that we have gradually come to grasp the nature of myriad other initially mysterious processes that occur in the world—the relation between heat and molecular motion, for example. The upshot, then, is just to say that consciousness may well turn out to be a biological phenomenon at its core, even if we cannot *presently* see how to construct a fully worked-out “realization theory” of mental phenomena in these terms.

Alternatively, the antecedent physicalist might equally exploit the fact that metaphysical physicalism is entirely compatible with an identity thesis of the mind. The central
claim here would be that, ontologically speaking, conscious states just are brain states, even though certain facts pertaining to the former cannot be gleaned through a priori reflection upon the latter (and vice versa). This, we might say, would constitute a kind of ontologically reductive but epistemically non-reductive account of the nature of phenomenal consciousness. More significantly, on this view, it is the pervasive dogma that consciousness arises or emerges from brain states that stands as the principal misconception in the entire debate. Indeed, the identity theorist further laments that this is a misconception that is tacitly nourished by both dualists and physicalists. This is because on the identity view there literally is no causal “nexus” that looms large and mysterious. So understood, the ongoing dualist-physicalist exchange about the proper form that any plausible “realization theory” of consciousness must take can, in effect, be written off as a kind of mutually sustained red herring; indeed, so much the worse for both sides of the fence. Or so, at least, the antecedent physicalist could argue.

4.2 Refrain: zombies, conceivable, and possibility

Before proceeding any further, we should pause briefly to address a very natural objection to everything that I have just said in the previous section. The objection goes as follows. By removing the EPL clause and the DL thesis from the physicalist story, one thereby edges steadily closer to endorsing a form of dualism, even if only tacitly. The claim here is that the physicalist’s retreat from the linguistic to the metaphysical variant of the doctrine at best blurs the distinction between physicalistic monism and Chalmers’ “naturalistic dualism,” and, at worst, threatens to dissolve the distinction entirely. So far as it goes, this claim surely does succeed in bringing to light an important point concerning some of the conceptual similarities between the two views on the table. The trouble for the dualist, however, is that as an objection to my proposal this clearly does not go very far. Again, this is because the antecedent physicalist does not assert the truth of metaphysical physicalism, but merely
states the view as a conceptually coherent *conjecture* about the way the world actually is, given our current knowledge base. Indeed, the antecedent physicalist also accords the very same honor to naturalistic dualism. This should not, however, be taken as an awkward admission, much less a significant concession from the antecedent physicalist. Antecedent physicalism, recall, is an essentially *defensive* posture, hence it is entirely consistent for the antecedent physicalist to accept the possibility that future evidence or argument may yet establish that dualism is true. Denying this as a conceptual possibility is no part of the antecedent physicalist’s project, nor is this something that one is somehow committed to nonetheless purely because one believes that the existing arguments for dualism are unsuccessful. But more importantly, despite some obvious similarities in terms of theoretical superstructure, the fundamental analytic distinction between the two views is quite clear, and thus very easy to state: if metaphysical physicalism is true, then, ontologically, the world is made up of just one kind of stuff; if naturalistic dualism is true, then the world is made up of two very different kinds of stuff. The crucial point here is that the antecedent physicalist acknowledges that both metaphysical physicalism and naturalistic dualism enjoy the status of what we might call *prima facie* conceivable speculative conjectures about the way the world is. Only one, however, is ideally conceivable, and hence only one is true.

A useful way to underscore the important distinction between *prima facie* and ideal conceivability would be to once again consider the notion of a zombie world. In truth, the relation between conceivability and possibility is a discussion that we have been destined to return to ever since the first section of the previous chapter, where we were forced to bracket the conceptual possibility of a “zombie world” as an as-yet open question. Let us now return to this question directly. The first thing to note is that with the metaphysical contra linguistic physicalism framework of Chapter Three in place, we are now in a much stronger position overall from which to engage the zombie question. In particular, we are now afforded an entirely different critical strategy than that which was pursued in Chapter Two. The fundamental strategy of Chapter Two, recall, was to argue against the inherent
epiphenomenalism of the zombie scenario, by suggesting that, contrary to Chalmers’ description, a zombie actually would not be functionally isomorphic to us from the third person point of view. As was conceded, however, this sort of critique, even if successful, is clearly insufficient by itself to overturn property dualism, since the dualist might yet retreat to the weak absent-qualia hypothesis, by asserting the conceptual coherence of a non-conscious being that, though behaviorally non-equivalent to us, is nevertheless microphysically identical to us in terms of its internal structural constitution. Now consider what was shown in Chapter Three. Here we saw that the truth of physicalism, properly understood, need not be equated with the impoverished thesis that everything physical must be fully describable in explicitly physicalistic terms. Even so, qualia may still be taken as ontologically physical phenomena insofar as they are most simply understood as physical properties of physical states. Nevertheless, given that qualia are only ever fully knowable in the first person, they will always elude the ken of direct scientific observation. To be sure, given the plausibility that Mary might form what I have termed “theoretical concept QR,” metaphysical physicalism openly concedes that qualia are at least partially describable in scientific terms. However, the reason that they are not fully describable in such terms is precisely because they are not the sort of thing that can be exhaustively investigated from the inherently third personal perspective of science. Now, if this much is granted then it follows naturally that we simply do not yet know whether zombies are ideally conceivable or are merely prima facie conceivable, because we cannot discount the possibility that in trying to imagine a world that lacks conscious experience entirely but is nevertheless physically identical to the real world, we are actually misdescribing what a zombie world would necessarily have to be. Granted, there does not appear to be anything prima facie incoherent in the notion of a hypothetical world populated by non-phenomenally conscious beings that look and perhaps even act just like we do. The crucial point, however, is that this seemingly clear and distinct description may already imply that such a world is physically non-identical to our own, because it is entirely plausible that we simply do not yet know enough about the
nature of phenomenal consciousness in our world to make a hard and fast determination one way or the other. Consider: we have seen that the metaphysical physicalist has an entirely plausible story about why subjective experience resists perspicuous description in the formal languages of the basic sciences; namely, because “perceptual capture is done in the first person.” We have also seen that the ontological conclusion that anything that resists such description should therefore be taken as somehow demonstrably nonphysical does not follow, and is in fact a non sequitur. However, from the conceptual possibility that metaphysical physicalism may actually be true, it clearly does follow that there could well be physical facts pertaining to human consciousness that are implicated in our explicitly physicalistic descriptions of the brain and the body, but are nevertheless not transparent in these terms. What this means, then, is that the standard description of a philosophical zombie may actually involve a formal contradiction after all, even if the relevant concepts do not seem internally incompatible from our current non-ideal perspective. On page 94 of The Conscious Mind, Chalmers describes a zombie as follows: “This creature is molecule for molecule identical to me, and identical in all the low-level properties postulated by a completed physics, but he lacks conscious experience entirely.” Now, if the metaphysical physicalist is right and qualia are physical in kind, then this description from Chalmers clearly does not hang together: that is, either my zombie twin is physically identical to me, in which case he is also, therein, my phenomenal doppelganger—and, hence, not actually a zombie—or he really is a zombie, but is therein necessarily physically distinct from me. Valerie Gray Hardcastle provides a nice summary of the problem: “the skeptics [she means the “consciousness-mysterians,” p.66] can’t really imagine possible worlds in which consciousness is not whatever we ultimately discover it to be because they aren’t imagining

---

82 John Searle (1997), for example, invokes this same general point when he suggests that the zombie argument is actually invalid. Searle claims that in trying to imagine non-conscious beings that are behaviourally indistinguishable and physically indiscernible from us, we invariably end up imagining, even if only tacitly, different laws of nature holding sway, and thereby violate the criterion of strict nomological equivalence across the two worlds (a criterion which is implied by the “physically indiscernible” stipulation). See his “An Exchange with David Chalmers”, reprinted in his book “The Mystery of Consciousness” (The New York Review of Books, 1997). See esp. pp. 146-148. Chalmers’ reply is also reprinted here.
consciousness in those cases (or, they aren’t imagining properly).”

To help bring this out, consider the imaginary philosopher’s beings known as shomies. Richard Brown provides an especially clear and concise definition of a shombie. More specifically, in his recent paper “Deprioritizing the A Priori Arguments Against Physicalism,” Brown presents an interesting “reverse thought experiment” and asks us to consider what might be dubbed the “shombie argument against dualism.”

A shombie is a creature that is micro-physically identical to me, has conscious experience, and is completely physical. Shombie pain is just as painful as my pain and shombie orgasms are every bit as pleasurable as mine are. My shombie twin and I have all the same experiences. The only difference, if it is a difference, is that shombie pain is completely physical. That doesn’t make it any different from the inside. What it is like for me to have a pain and what it is like for my shombie twin to have a pain are identical in all respects. We have stipulated that shombie pain is just like my pain in every respect (qualitatively) and that my shombie twin is a complete micro-physical duplicate of me in a world that is stipulated to lack nonphysical properties of any kind and that this is all there is to a shombie. The shombie is NOT a zombie. A zombie lacks phenomenal consciousness; a shombie doesn’t. The qualitative does, therefore, logically supervene on the physical and dualism is false. Zombies are metaphysically impossible.

We can formalize the shombie argument as below: where P is a complete physical description which includes a “that’s all” phrase and Q is some qualitative fact about me, say that I am now seeing green.

1. P and Q is conceivable.
2. If (P and Q) is conceivable, then (P and Q) is possible.
3. If (P and Q) is possible then dualism is false.
4. Therefore dualism if false.

The point of this reverse thought experiment exercise from Brown, in the first instance, is to draw out the following natural complaint from the dualist, which can be summarized as follows. Consider: the so-called “shombie argument,” such as it is, is plainly not compelling as an argument against dualism since its physicalist architect has merely stipulated that conscious experience is to be understood as a physical phenomenon. In other words, rather than to argue for this conclusion directly, the shombie advocate is illicitly trying to smuggle it in at the outset, by passing it off as if it were actually a premise. As any first-year

philosophy student knows, it is quite clear that “conscious experience is completely physical” just will not do as a premise in an argument that, after all, aims to establish that conscious experience logically supervenes on, or is identical to, the physical, since this is precisely what is at issue in the debate to begin with! In sum, the complaint from the dualist is that the shombie argument is straightforwardly question-begging; an attempt to win an argument by definitional fiat, nothing more. From here Brown invites the reader to stand back a few paces and to consider the overall fairness of the dualist’s complaint, particularly in light of the fact that, as he rightly suggests, a physicalist might equally object in exactly the same terms to the way the zombie argument is constructed. To see that this is so, consider this passage from the opening page of Chapter 4 of The Conscious Mind, where Chalmers provides the following formal statement of the zombie argument.

1. In our world, there are conscious experiences.
2. There is a logically possible world physically identical to ours, in which the positive facts about consciousness in our world do not hold.
3. Therefore, facts about consciousness are further facts about our world, over and above the physical facts.
4. So materialism is false.85

Both the shombie and zombie arguments obviously share Chalmers’ first premise. The difference is that the dualist zombie advocate does not accept premise 1 of the shombie argument; while the physicalist shombie advocate does not accept premise 2 of the zombie argument, which, like our imaginary dualist critic, she insists is clearly question-begging. Now, it is clear enough that someone must be begging the question here; what is not clear, however, is which side is the guilty party. As Brown puts it, I think rightly, the problem is that “both sets of arguments are exactly parallel [in terms of their logical structure] so there is no good a priori reason to say who is doing the begging.”86 To be sure, both arguments seem quite conceivable, and in that sense, we might say, both are equally convincing—at least in the narrow sense that both seem “negatively conceivable” in that neither involves

86Brown, “Deprioritizing the A Priori Arguments Against Physicalism,” p. 11.
any explicit internal contradiction. Of course, the mere fact that from our current perspective both arguments appear equally intuitively plausible to us obviously does not mean that, as metaphysical theses about the way the world actually is, neither is any more or less likely than the other to turn out true. Objectively speaking, there is of course a truth of the matter here, in that either shombies or zombies, but not both, are ideally or “positively conceivable” and hence metaphysically possible. We just don’t know which one is which. Notice also the striking parity between our dilemma here and that of Mary’s plight in the Nida-Rümelin room. For Mary, recall, $QR$ was either $Q\text{-this}$ or $Q\text{-that}$, but there were no $a\ priori$ reasons that told in favour of one possibility over the other. Prior to her release, then, that “$QR$ is $Q\text{-this}$ is true” was exactly equally plausible as that “$QR$ is $Q\text{-that}$ is true.” One way to put the point would be to say that both identity statements were equal “epistemic possibilities” for Mary, though only one was metaphysically possible. Wisely, Mary withheld final judgement on the matter until such time as she was able to empirically ascertain which hypothesis described the actual state of affairs in the world. In effect, Brown counsels the same general strategy about how physicalists should respond to the zombie argument. In sum, a zombie world surely is $prima\ facie$ conceivable; but so is a shombie world. We simply do not yet know which of the two is ideally conceivable and which is merely intuitively plausible. It does seem, however, that in taking seriously the moral of the reverse thought experiment of the shombie scenario, we have found good reason to concede that no amount of armchair $a\ priori$ reasoning can be relied upon to settle the matter. Hence the sensible course forward, at least for now, is to refrain from making any concrete “all-in” claims about either scenario, and to instead wait on further $a\ posteriori$ discoveries about consciousness and its place in the natural world. Nevertheless, given all that we have seen regarding how poorly the standard anti-physicalist arguments fare on close inspection, the commonsense plausibility of metaphysical physicalism is striking. Thus, I conclude, once physicalism has been shorn of the various doctrines that it need not and should not endorse, it becomes clear that the case for dualism has not been made.
Bibliography


Horgan, Terrence. “Jackson on Physical Information and Qualia”, Philosophical Quarterly 32: 1984, pp. 147–152.


71