9-2019

Mentality and Fundamentality

Christopher D. Brown
The Graduate Center, City University of New York

How does access to this work benefit you? Let us know!
Follow this and additional works at: https://academicworks.cuny.edu(gc_etds
Part of the Metaphysics Commons, and the Philosophy of Mind Commons

Recommended Citation
https://academicworks.cuny.edu(gc_etds/3299

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
MENTALITY AND FUNDAMENTALITY

By

CHRISTOPHER DEVLIN BROWN

A dissertation submitted to the Graduate Faculty in philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2019
Mentality and Fundamentality

By

Christopher Devlin Brown

This manuscript has been read and accepted for the Graduate Faculty in philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date

Graham Priest
Chair of Examining Committee

Date

Nicholas Pappas
Executive Officer

Supervisory Committee:

David Papineau
Barbara Montero
Torin Alter
Susan Schneider
Graham Priest

THE CITY UNIVERSITY OF NEW YORK
ABSTRACT

Mentality and Fundamentality

By

Christopher Devlin Brown

Advisor: David Papineau

Extant well-considered problems with physicalism primarily come from two sources: philosophers of mind arguing that subjective experience does not fit into a physicalist world-picture, and metaphysicians trying to figure out the particular commitments of the view. I examine the thesis of physicalism in order to produce a clearer notion of the physical and to help straighten out physicalism’s entailments, while simultaneously providing a strategy for physicalists to sidestep well known anti-physicalist arguments concerning consciousness. This involves both a critical and a positive effort: on the critical side, I expose an issue with a popular way of understanding physicalism called “via negativa” physicalism, which is the view that ‘physical’ should be understood to mean ‘not fundamentally mental’. The positive project has two components. One part defends physicalism from the ever-loomong threat of the scientific insolubility of phenomenal consciousness by fleshing out a version of Russellian monism called “Russellian physicalism”, which is the view that the structural/dispositional properties described by fundamental physics have inscrutable role-fillers that are not directly revealed through scientific inquiry. The other part of the positive project consists in examining the existentially relevant consequences of physicalism, which are the consequences that have a bearing on whether we ought to continue to live. Toward this end, I determine whether such things as robust free will, God, ultimate purpose and an immortal soul are consistent with the most popular versions of physicalism.
Table of Contents

I. Introduction
   1. Summary
   2. Historical Preliminaries

II. Problems for *Via Negativa* Physicalism
   1.1 Infinitely descending mentality a physical world
   1.2 Physicalism and infinite descents
   1.3 A compositional account of mentality
   1.4 Quasi-brains all the way down
   1.5 Conclusion
   2.1 Fundamental mentality in physical worlds
   2.2 Universe-sized Minds in Priority Monist Worlds
   2.3 Objections
   2.4 Conclusion

III. The Hope and Horror of Physicalism
   1. Two Notions of the Physical
   2.1 Free will
   2.2 Free Will and *Via Negativa*
   2.3 Free Will and Theory-Based Physicalism
   3.1 God
   3.2 God and *Via Negativa*
   3.3 God and Theory-Based Physicalism
   4.1 Immortal Soul
   4.2 *Via Negativa* and Immortal Soul
   4.3 Theory-Based Physicalism and Immortal Soul
   5.1 Ultimate Purpose
   5.2 *Via Negativa* and Ultimate Purpose
5.3 Theory-Based Physicalism and Ultimate Purpose

6.1 Karma

6.2 *Via Negativa* and Karma

6.3 Theory-Based Physicalism and Karma

7. Epicurean Freedom and Scientific Imperialism

8. Conclusion

IV. A Minimally Mysterian Solution to the Mind-Body Problem

1. What is Russellian Physicalism?

2. McGinn’s Mysterianism

3. The Phenomenal Concepts Strategy

4. Conclusion

V. Bibliography
I. Introduction

1. Summary

Physicalism—roughly, the view that there is nothing over and above the physical—has nearly achieved the status of orthodoxy in analytic philosophy, serving as a backdrop against which diverse topics such as the nature of representational content (Fodor 1990), free will (Pereboom 2014) and the self (Parfit 1984) are discussed. In philosophy of mind, arguments concerning physicalism have occupied a central role following Jackson’s (1982) articulation of the knowledge argument and Chalmers’ (1996) articulation of the conceivability argument, both of which conclude that physicalism cannot accommodate the existence of subjective experience. However, despite the massive importance of physicalism across the field, there is no consensus on what the view specifically means. For instance, some think physicalism has to do with the dependence of all things on the entities and properties described by physics, while others think this physics-based approach is a dead end, and instead urge that physicalism has more to do with the sort of mentality that exists (or doesn’t exist) in an entirely physical world.

My primary goal in this dissertation is to work toward the proper articulation of physicalism, and figure out what its consequences are. I start by critiquing a way of understanding physicalism called “via negativa physicalism”, which says that physicalism is true if everything ultimately depends on things that are not mental. Under this view, fundamental mentality of any sort renders physicalism false. I argue that via negativa wrongly rules that physicalism is false in various classes of possible worlds in which physicalism plausibly ought to be true. These are worlds with either an infinite descent of physically-acceptable mentality in an infinitely decomposable world, or physically-acceptable mentality at the level of the whole cosmos in a priority monist world.

The general idea behind via negativa physicalism is right—nearly everyone agrees that the existence of certain sorts of mentality is sufficient to make physicalism false. The problem is that via negativa as typically articulated does not offer a nuanced enough distinction between physically acceptable versus physically unacceptable mentality. Thus, my criticism ought to motivate via negativa physicalists to find a more acceptable account of the physical, and specifically one which distinguishes
between physically acceptable mentality and physically unacceptable mentality in a more sophisticated way than via a simple appeal to fundamental mentality.

The third section of the dissertation considers physicalism as an existential condition, which I define as a condition that has a bearing on whether we ought to continue to live or not. Upon inspection, it turns out that different versions of physicalism have different existential consequences. For example, any plausible theory-based definition is consistent with the existence of an ultimate purpose for the universe or individual humans, but the \textit{via negativa} definition is inconsistent with the existence of ultimate purpose. I will show that \textit{via negativa} physicalism is the existentially more radical sort of physicalism—terrifying for those in need of some metaphysical reassurance that humanity is somehow metaphysically special, and emboldening to those who think humanity ought to cast off the superstitious myths of the past.

The final section defends a version of physicalism called “Russellian physicalism” as the best answer to the mind-body problem. This is the view that there are scientifically inscrutable, categorical realizers for the structural and dispositional properties described by physics. These categorical realizers form the dependence base for higher-level inscrutable, categorical properties, amongst which are the properties that constitute subjective experience. This view provides satisfying answers to all the typical antiphysicalist challenges offered by nonphysicalists, and these answers do not require making the fundamental categorical properties essentially mental or tied to mentality in a problematic way.

It turns out that this Russellian physicalism is consistent with \textit{via negativa} physicalism, but not with theory-based physicalism. This is because the view explains subjective experience without positing fundamental mental properties, but additionally requires that there be properties outside the purview of science. Given the attractiveness of Russellian physicalism, it is my view that we need to face the more extreme existential consequences of \textit{via negativa} physicalism.

Some of the material in this dissertation is pulled from or closely related to work that I have published or which is currently under review. Specifically, the first half of section two, which criticizes \textit{via negativa} physicalism, is largely from my 2017 Erkenntnis paper, \textit{Minds Within Minds: An Infinite Descent of Mentality in a Physical World}. The second half of that section is from a paper that is under review. Everything in section three, on physicalism as an existential condition, is original to this dissertation. The
final section, on Russellian physicalism, is from a paper that is under review, and is related to work in my 2017 Journal of Consciousness Studies paper, *A Properly Physical Russellian Physicalism*.

2. Historical Preliminaries

Before getting any further into the fine details, allow me to first tell a story about how physicalism arose and came into prominence, as well as the historical relationship between abstract metaphysics generally and humanity’s various existential concerns. Much of what I say here is just a combination of strands of thought found in David Papineau’s (2001) *The Rise of Physicalism*, Jaegwon Kim’s (1998) *Mind in a Physical World*, Barbara Montero’s (1999) *The Body Problem*, and Tim Crane and David Mellor’s (1990) *There Is No Question of Physicalism*. However, I do not agree with Crane and Mellor’s conclusion that physicalism must be an empty doctrine, and have reservations about Papineau’s thesis that twentieth century scientists acquired sufficient empirical evidence to prove the view—but I am getting ahead of myself.

Starting at the beginning: Thales, of first-Greek-philosopher fame, argued that everything is water. So far as I can tell (from the little we know about him, which comes entirely from secondary sources like Aristotle), he did not think that everything literally is water in the ordinary sense. That would be absurd. Rather, he was asserting that everything is *fundamentally* water: that the most basic constituents of the world are watery. By contemporary standards, the view is silly and drastically undermotivated, but it nonetheless represents a vital milestone. The significance consists not in the view itself, but rather in its general structure and the approach that led to it. Thales was positing a novel thesis about the underlying fundamental nature of reality, and he did not appeal to tradition or divine inspiration to support his thesis. Rather, his metaphysics was grounded in empirically-based reasons: according to Aristotle, Thales noticed and tried to explain that all living beings get nourishment from moisture.

Thales, of course, opened the floodgate for a torrent of philosophical theorizing about the fundamental nature of reality—which I mostly won’t get into here. What is relevant is that these philosophers took the views they were promulgating to have not just scientific or abstract significance, but additionally to have a great deal of personal significance. For the ancient Greeks, existential cares were intimately tied up with their metaphysical visions, often quite explicitly.
For instance, Plato thought that both individual and collective good depend on realizing an appropriate relationship to the perfect and eternal forms. Individuals and governments can be ranked into hierarchies depending on their relationships to the forms, with the bottom being those persons and societies that are not informed by the forms at all, and the top consisting of Platonic philosophers and societies run by Platonic philosophers who have intimate knowledge of the forms. Offering a quite different metaphysical and existential perspective, Epicurus thought that the truth of atomistic materialism reveals that we do not have immortal souls which can transcend the body after death, and so we need not fear what would happen to us in an afterlife. Epicurus found existential comfort in the transience of human existence—nearly an exact inversion of Plato, who deemed highest value to reside in the eternal, and espoused endless transmigration of the soul from body to heaven and back again, cycling eternally, as an enlightened and hopeful doctrine.

Tying existential cares to a metaphysical project continued well past the Greek period. A random sample: the Stoics believed that the best way to live is in accord with their idea of nature; many Christian philosophers held that the best life is one in close relationship to the divinity; Hegel thought that human good is determined by universal consciousness rising to higher levels of self-consciousness; and Schopenhauer thought that all of our near-inescapable sufferings result from the activity of the will of the world, which wretches in the dark at the bottom of all things. No deep philosophical exegesis beyond this cursory surface-level look is necessary to reveal that existential interests have played a large role in the metaphysical theorizing of many of our most significant philosophers. But what about my principal subject, the all-too-contemporary thesis called ‘physicalism’?

How physicalism arose and developed will form the basic outline for a picture of its existential import. Physicalism can be understood as a view that emerged from two related threads of discourse. One thread is philosophical and scientific discussion concerning the relationship between the mind and the brain—referred to in philosophical literature as the “mind-body problem”. The other thread consists of philosophers attempting to articulate the appropriate metaphysics of science. These two threads cross at several points, and in fact have become quite tangled.

As everyone who has studied any philosophy of mind knows, the contemporary mind-body problem more-or-less started with Descartes (1641/1996), who posited two sorts of fundamental
substances: the thinking and feeling mental, and the extended material, which interact via the immaterial mind’s ability to change the direction of the material components of the brain. The nature of this interaction, though, is quite mysterious: how can something which is essentially immaterial cause changes to an extended, material substance? According to Leibniz (1714/1991), Descartes’ answer rested on an incomplete understanding of physics: his conservation of motion law allows change of direction of extended bodies so long as the total amount of motion is conserved, such that the mind can cause bodily behavior by changing the directions of extended things moving in the brain, yet does not change the total amount of motion in that system.¹

The subsequent physics of Newton and Leibniz posited conservation laws that disallow changing the direction of an object without the addition of energy, and necessitated the need for a new explanation of the relation between mind and body. Importantly, however, Leibniz’s physics requires contact-based interaction between objects, whereas Newton posited force fields that allow for non-contact-based interaction—and notably allow for special mental forces. Thus, Leibniz’s (instantly outdated) physics elicited epicycles on Cartesian dualism, as in Leibniz’s own pre-established harmony and Malebranche’s occasionalism, whereas Newtonian physics allows for a uniquely mental force of nature—which apparently made interactionist mind-body dualism a scientifically respectable option.

However, according to David Papineau (2001), further scientific progress in the twentieth century and onward has provided evidence against the positing of special mental forces, and thus there seem to now be good empirical grounds to rule out interactive Cartesian-esque dualism. Since both epiphenomenalism (which says that the mental is not causally efficacious) and overdetermination (which says that physical effects are overdetermined by both mental and physical causes) are unpalatable options, materialism emerged as the apparent best solution to the interaction problem.²

¹ This is likely due to Descartes thinking of the behavior of material bodies in terms of the scalar concept of speed rather than the vectorial concept of velocity. The total motion of a system, determined by the speeds and (for Descartes) sizes of the parts of the system, can remain constant even if parts of the system change direction. Not so for velocities: change of direction requires a change of velocity. The given reading of Descartes—that his good-for-the-time but poor-for-now physics grounds his tentative answer to the interaction problem—is somewhat contentious, and largely based on an interpretation from Leibniz. However, see Peter McLaughlin (1993) for a defense of this exegesis.

² Of course, this gloss on the history of the mind-body problem overlooks many major developments in philosophy of mind and philosophy more generally, perhaps most notably the positivist period that dominated analytic philosophy in the first half of the twentieth century. While metaphysics was explicitly
mind and matter, it seems that there is no issue with interaction: the material mind can affect other parts of the material world in roughly the same way that any other material things interact.

Additionally, this view seemed to be the proper metaphysics of science: one might think that our best science only posits matter and laws governing matter, so any metaphysics positing nonmaterial entities would exceed our scientific reach. However, materialism as an a priori project of specifying the nature of the fundamental world turned out to be—contrary to the scientific spirit of materialism—an antiscientific philosophical endeavour. This is because a priori conceptions of matter—concocted from the armchair rather than from the laboratory—attributed many properties to fundamental entities that our best physics now denies. For instance, 17th century materialists’ conception of matter was of something “solid, inert, impenetrable and conserved, and to interact deterministically and only on contact” (Crane and Mellor 1990, 186). So materialism was a misguided attempt at articulating the proper metaphysics of science. This does not mean materialism was a complete failure: it birthed an apparently more viable successor, the view called “physicalism”.

Physicalism attempts to avoid the error of materialism by leaving somewhat open exactly what the fundamental level of reality is like. Originally, physicalism did this by understanding ‘physical’ solely in terms of physics: something is fundamental and physical if it is a posit of fundamental physics. Thus, philosophers thought to leave the task of determining what the fundamental level of reality is precisely like to the physicists, rather than characterizing reality from the armchair.

However, Hempel (1969) pointed out a devastating ambiguity in this understanding of ‘physical’: if ‘physics’ refers to current physics, then physicalism is trivially false, since we know that current physics is not complete. If ‘physics’ refers to future or complete physics, then physicalism is either vague to the point of meaningless, since we don’t know what future physics will be like, or trivially true, since a complete physics is trivially a physics of everything that exists, including paradigmatically nonphysical things like Cartesian souls and God. This problem has constituted the strongest motivation to find a more adequate notion of the physical in a definition of physicalism.

anathema for these philosophers, the default position on the nature of the mind was one or another sort of behaviorism, which might be interpreted as a materialist or materialist-leaning view. I’m not sure how to perfectly square positivist behaviorism with their apparent commitment to phenomenalist idealism of some sort, but here I’m not specifically interested in this period in the history of analytic philosophy, so the gloss seems appropriate.
Largely independent of this discussion in metaphysics, work progressed in philosophy of mind toward establishing a theory of the mind that would be consistent with the newly dominant materialism/physicalism, and which would satisfy an expanding list of various desiderata. Behaviorism—a view closely tied to anti-metaphysical logical positivism, which says that non-verifiable (and hence meaningless!) talk about inner mental states cannot be part of the content of a properly scientific theory of the mind (Hempel 1949, Ryle 1949)—was the dominant theory for the first half of the twentieth century, and in its strongest form identified mental states with behavioral dispositions. The view died for various good reasons, and initially gave way to type-identity theory (Smart 1959, Feigl 1958, Place 1956), which identifies mental properties with neural properties. Identity theory was quickly superseded by functionalism (Putnam 1967), which says that mental properties are identical to causally-defined functional properties.

The most serious difficulty these physicalist theories of mind now face is the charge that no such theory can adequately account for subjective experience. This is the now-well-known “hard problem” of consciousness (Chalmers 1995): it is difficult to reconcile the existence of rich subjective experience with a purely physicalist account of nature. It can be supposed that the brain is a sort of naturally evolved, entirely physical machine—a rather unusual, wet, chaotic and enormously complex machine, but nonetheless a machine. If we start from a simple environment-registering and behavior-modifying machine like a thermostat, it it hard to image how adding more complexity, or even introducing biologically-based mechanisms, could generate the “what it is like” of subjective experience.

Philosophers such as Frank Jackson (1982) and David Chalmers (1996) have articulated these intuitions into more precise arguments, which have various modal and other suppositions built into them, but I think this idea—that no mere physical machine, which is nothing more than a dynamic arrangement of lower-level unfeeling physical components, could really feel like we do—is at the heart of the issue. The via negativa account of physicalism—which emerged out of a perceived insufficiency of theory-based physicalism to address Hempel's dilemma—endeavors to explain the self-identification of philosophers

---

3 This is not to suggest that the hard problem is the only source of criticism for any given physicalist theory of mind. For instance, any such theory ought to comport to the available empirical evidence, which is a source of more specific criticism of particular theories, not to mention a host of other considerations that apply to appropriate theory selection. However, I take it that the hard problem is more influential and wide-ranging than other, more specific objections to particular views.
working on this problem as ‘physicalist’ or ‘nonphysicalist’ relative to their commitment to fundamental mentality. *Via negativa* physicalism says that physicalism is true if there is nothing which is both fundamental and mental. God and Cartesian souls seem to violate this criterion—since they are generally defined as fundamentally mental phenomena—whereas brains and computer systems do not—since their mentality ultimately depends on the nonmental properties of their parts.

This brings my narrative of the development of physicalism up to the present. We’ve seen that physicalism is the view that replaced materialism as the appropriate metaphysics of science, which would simultaneously provide a satisfying and plausible answer to the problem of how mind and body interact. There are two popular ways of understanding physicalism: theory-based physicalism and *via negativa* physicalism. I indicated that many—perhaps all—of the significant metaphysical accounts from the past bear relationships to those philosophers’ existential concerns. What is the relationship between physicalism and our existential concerns? Does physicalism also have some relationship to our personal hopes, cares and fears? What, if anything, is at stake if physicalism is true or false? Before I can answer these questions in section three, some ground clearing is in order.
II. Problems for via negativa physicalism

The metaphysical view called ‘physicalism’ asserts that there is nothing over and above the physical—if physicalism is true, then everything that exists is identical to, constituted by, or entirely dependent upon physical things. This view has become nearly orthodox in contemporary analytic philosophy, often operating as an assumed background condition against which other discussions occur. Despite its massive popularity and influence, physicalism itself is not well understood. This is because the ‘physical’ in ‘nothing over and above the physical’ is underdefined, with different precisifications yielding different sorts of physicalism—and as yet, there is no consensus on the right way to refine the definition.

A common way to understand the physical is as whatever is referred to by our best physics (Hellman and Thompson 1975; Smart 1978; Hellman 1985; Poland 1994; Melnyk 1997, 2007; Stoljar 2001). Unfortunately, this theory-based approach faces Hempel’s (1949, 1969) well-known dilemma: ‘physics’ is ambiguous, and can either refer to current physics or future physics. If ‘physics’ refers to current physics, a problem arises: there are yet-to-be-discovered entities and properties which are outside of the scope of current physics. Because current physics does not refer to these undiscovered things, their existence makes current-physics physicalism false. However, if ‘physics’ refers to future physics, then physicalism is vague to the point of meaninglessness, since we do not know what the content of future physics will be.

This dilemma from Hempel has constituted the strongest motivation for philosophers to come up with a more adequate notion of the physical. Perhaps most notable is the via negativa account (Smith 1993, Papineau and Spurrett 1999, Levine 2001, Montero 2005, Montero and Papineau 2005, Worley 2006), which in its most basic form says that ‘physical’ means ‘not fundamentally mental’. I follow Jonathan Schaffer (2010) in understanding ‘fundamental’ to mean ‘depends on nothing else’. Properties that bear metaphysical dependence relations between each other—as opposed to causal dependence relations—follow a hierarchical ordering from less to more fundamental.

Something is fundamentally mental if it is mental and does not depend on anything that is more fundamental—mentality of this sort is at the bottom of nature, so to speak. Cartesian souls, assuming they are fundamentally mental things, count as non-physical under this understanding of ‘physical’;


whereas subatomic particles, if they are both fundamental and non-mental, are physical. Objects such as brains, which seem to be entirely composed out of non-mental parts, with properties that ultimately depend only on non-mental properties, can possess non-fundamental mentality without making via negativa physicalism false.

Not everyone agrees that the via negativa approach suffices to explain what it takes for something to be physical. For instance, Seth Crook and Carl Gillett (2001) and Jessica Wilson (2006) assert that this is a necessary condition on being physical, but not sufficient—they agree that physicalists should accept a “No Fundamental Mentality” constraint, but think something more is required to flesh out the notion of physicality. However, nearly everyone agrees that a No Fundamental Mentality constraint is, at minimum, a necessary condition for any version of physicalism: if anything exists in a world which is both fundamental and mental, then physicalism is rendered false in that world, regardless of whatever else physicalism requires.

Unfortunately for those who would like to use via negativa or No Fundamental Mentality to define or help define physicalism, closer inspection reveals that these principles must be rejected. This is because deeming that physicalism is false in any world which contains fundamental mentality leads to the wrong ruling on certain possible worlds that physicalists have hitherto largely neglected. These are (i) worlds which are have no fundamental properties, and (ii) worlds in which priority monism is true.

The traditional view of metaphysical dependence says that all properties in a world ultimately depend on the properties of the most mereologically basic constituents of that world—I’ll call this traditional view “priority partism”. Worlds which are infinitely decomposable contain no mereologically basic components—and so no fundamental properties, if priority partism is accepted. In worlds where priority monism is true, the fundamental properties are the properties of the whole cosmos, rather than the smallest bits of the universe. Both of these types of worlds can contain mental properties which via negativa must say are physically unacceptable. Yet, when we look at the details, it turns out that certain of these mental properties ought to count as physically acceptable by any reasonable standard. So via negativa and No Fundamental Mentality must be rejected or radically revised.

What follows is organized into two main parts, which are taken from two different papers. The first major part (numbered 1.1, 1.2, …) discusses an infinite descent of physically acceptable mental
properties in an infinitely decomposable world. The version of via negativa that has been developed to accommodate such infinitely decomposable worlds wrongly rules that physicalism must be false in such a world. The second major part (numbered 2.1, 2.2, ...) discusses physically acceptable mental properties realized at the level of the whole of nature in priority monist worlds, which trivially means they are fundamental under priority monism. Via negativa and No Fundamental Mentality straightforwardly make the wrong ruling on these worlds.

1.1. Infinitely descending mentality a physical world

Physicalism is typically thought of as the thesis that everything depends upon a fundamental physical level. In Barry Loewer’s words, physicalism is true if and only if “all facts obtain in virtue of the distribution of the fundamental [physical] entities and properties” (Loewer, 2001, p. 2). On this reading of physicalism, the world is composed of levels with hierarchical dependence relations between them; put more simply, physicalism says that the way that things are at a higher level is determined by how things are at lower levels. For instance, the macro-level property of fluidity that belongs to water depends upon chemical properties of water molecules, and those properties depend in turn upon atomic properties of hydrogen and oxygen atoms, and so on until at some point one arrives at the fundamental level, which is the level upon which everything else depends.

That is the standard view of physicalism; however, it has an infrequently noted and arguably unacceptable consequence. Namely, as discussed by Montero (2006), Brown and Ladyman (2009), Schaffer (2003), and Nagasawa (2012), it rules physicalism as false in worlds that have no fundamental level. For example, the standard formulation makes physicalism come out false in worlds in which things like fermions and bosons, which physicists currently take to be fundamental, are instead composed of other lower-level things, and the lower-level components of the fermions and bosons are themselves composed of even lower-level things, and so on ad finitum. Although there are people who revile physicalism, they don’t revile it for this reason. Rather, physicalists and nonphysicalists alike think that if

---

4I’m using the term ‘thing’ loosely, and not to mean ‘object’ or ‘entity’. I want to remain neutral on whether levels are composed of properties, entities, events, states, facts, etc. These are important distinctions, but here I do not think it will matter, and I will sometimes shift between these different ways of thinking of levels. My natural inclination, which I think is a straightforward approach and is used by Kim (1993), is to think of levels in terms of mereological relations between entities, with properties realized by entities at a given level.
physicalism turns out to be false, it will be because there are non-physical things in the world, and not because we live in a world that has no fundamental level.

Is the mere conceivable of a world with an infinite descent of levels something that those who aim to formulate a viable version of physicalism need to worry about? After all, many physicists today take the Planck length, which is the theoretical shortest measurable length as determined by the uncertainty principle, to be the smallest spatial value that is meaningful to discuss, as notions of space and length seem to be no longer applicable at conceivably smaller distances. While this is what current physics says about the smallest scale at which physical things can possibly exist in our world, our formulation of physicalism should not rest on this specific claim in physics: it is epistemically possible that the current view about Planck length is false, and if false that should not affect the truth or falsity of physicalism. Further, we should be seeking to have a notion of physicalism applicable to non-actual worlds, such as the infinitely decomposable worlds just described, regardless of the how the actual world is structured.

Can physicalism be revised so that it does not refer to a fundamental level? The only attempts to do this—most significantly in Montero (2006), Brown and Ladyman (2009) and Schaffer (2015)—designate a divide in the world between a subvening set of non-mental bottom levels and a supervening set of mental top levels. These formulations say that physicalism is false in worlds with an infinite descent of the mental, that is, worlds which are mental all the way down. Yet, as I will argue, physicalism can be true even in worlds with infinite descents of mentality, and thus such formulations of physicalism ought to be rejected. In other words, we ought to reject any formulation of physicalism that entails that physicalism is false in worlds with an infinite descent of mental properties. Not only that: because those formulations are the only attempts so far to specify a thesis of physicalism applicable to infinitely decomposable worlds, and assuming that it is reasonable to see such worlds as possible, this means we are left with no satisfactory dependence-based account of physicalism.

I will get to this conclusion by closely looking at the view that has been developed to make physicalism consistent with worlds that have infinite descents. After this, a general account of mentality will be developed which is consistent with various physicalism-friendly theories of mentality. Finally I will use that account of mentality to describe a world that is completely mental at all of its infinite levels of decomposition, yet in which physicalism might plausibly be true.
1.2. Physicalism and infinite descents

How can we formulate physicalism so as to account for the possibility of a world that has no fundamental level? Here, thus far, is the only approach: rather than stating physicalism as the thesis that everything depends or supervenes on an entirely physical fundamental level, take physicalism to be the thesis that everything depends or supervenes on phenomena that themselves depend only on physical phenomena.

On this approach, one assumes that there is some line that naturally distinguishes the fundamentally physical set of infinite bottom levels from the set of higher levels that depend upon them. The world is like a layer cake, similar to how it is in the hierarchical model that includes a fundamental level, but it is a cake with an infinite number of lower layers instead of a single bottom layer, and some significant distinction between the lower levels and the upper levels. To further this imperfect but hopefully helpful analogy, think of all the layers below some point on the infinite towering cake as dense chocolate, and they support foamy and insubstantial whipped cream layers above. The infinite descent of dense chocolate is the narrowly physical base for the broadly physical light whipped cream layers.\(^5\)

On Jonathan Schaffer’s view in Is There a Fundamental Level? (2003), this picture is not only imperfect but is of no use at all since, as he sees it there, "(p)hysicalism is an irreparably foundationalist doctrine because it requires an ontologically motivated distinction between the primary physical base, and that which, like the mental, is derivative"(2003, pg. 10).\(^6\) Further, "(t)he fundamental level provided such a distinction, by introducing a discontinuity in nature where the pattern of division was broken"(2003, pg. 10). The distinction between what is fundamental and everything else is, on this view, sufficiently motivated to stand as the dividing line between physical base properties and those that depend on them, while any non-fundamental level we might choose as our line would be arbitrary. For instance, in a world without a fundamental level, we might choose the atomic level as the dividing line such that if everything above that level depends on everything below it, physicalism is true. But why should we choose the

---

\(^5\) Brown and Ladyman (2009) correctly point out that such a layer cake model of the physical world is far too simple—the world is considerably messier than that, and if the metaphor were accurate at all the cake would be wildly uneven and have many sections not supporting any whipped cream. Still, I think the metaphor is useful as a rough visual illustration of levels of nature. A further issue: some philosophers, e.g. Dupré (1993), are highly skeptical of any levels-based metaphysical account. While such skepticism may be warranted, here I am responding to literature that takes a levels-based model as a starting point, and thus am assuming such a model.

\(^6\)As noted above, Schaffer has more recently changed his mind on the relation between physicalism and infinitely decomposable worlds, and now advocates Montero’s (2006) view, but I am using his earlier position for dialectical purposes.
atomic rather than the chemical or the biological, or even the psychological? Returning to the cake analogy, the chocolate/whipped cream divide in a simple two layer cake is significant, but—on Shaffer’s view—once we have an infinite number of layers, no divide is any more significant than any other. So Schaffer thinks physicalism is false in infinitely decomposable worlds. But must it be?

Montero (2006) contends that physicalism can still be true in infinitely decomposable worlds, and that there is a plausible non-arbitrary point at which we can demarcate the subvenience base from what supervenes on it. Schaffer even points to it in the quotation above when he refers to the derivative mental that is distinct from the fundamental physical: there is a divide in nature between the mental and the non-mental. For Robin Brown and James Ladyman (2009) this division is what unites early materialist and later physicalist world-views. On this account of physicalism, the physical subvenience base is the infinite set of all the non-mental levels below the lowest mental level. Supposing mentality exists only at the person level in the actual world, then the non-mental neural, chemical, atomic, subatomic, and so forth levels below the level of persons form the subvenience base. These philosophers assert that physicalism can be true in infinitely decomposable worlds so long as there is no mentality below some level. And an infinitely decomposable world that is mental all the way down, or at which one can always find a lower mental level, is a world in which physicalism is false.

The solution to the infinite descent problem for physicalism offered by Montero is consistent with a broader stance on how to define physicalism, labeled “via negativa physicalism.” This view, which seems to be growing in popularity, defines physicalism in terms of what it is not, delineating physical worlds as those worlds that do not instantiate things like ghosts, gods, miraculous events, souls or human-independent moral laws. This is the strategy adopted, in one form or another, by Montero (2005), Spurrett and Papineau (1999), Levine (2001), Worley (2006) and Montero and Papineau (2005). Montero’s response to infinite descent—which is a rather sophisticated version of this view that describes which levels, designated in terms of dependence relations to each other, will make physicalism false by instantiating mentality—cannot be invoked without also advocating via negativa physicalism, since it rules which infinitely decomposable worlds are physical by which properties are not realized infinitely below some level.
Could a non-*via-negativa* notion of physicalism be modified to accommodate worlds without a fundamental level? Perhaps Stoljar’s (2001a) theory-based account of physicalism, which roughly says that physicalism is true if and only if everything fundamental that exists is specified by physics, could be amended to include the claim that physicalism is false if there is an infinite descent below some level of entities or properties that are not specified by physics. However, this seems to give Hempel’s (1969) well-known dilemma against theory-based formulations of physicalism more bite in any worlds without a fundamental level: if ‘physics’ is specified to mean current physics, then physicalism is radically incomplete and thus false. If ‘physics’ is taken to mean a future and complete physics, then physicalism encompasses so many more things than we are familiar with that the thesis becomes vague to the point of meaninglessness. Regardless, the via negativa criterion is generally taken to be at least necessary for physicalism (as in Wilson 2006), so the argument I present should be a challenge to other conceptions of physicalism even if they are able to accommodate infinitely decomposable worlds.

Montero’s solution would be a great way to rescue physicalism. The only problem is that it doesn't work. The reason it doesn't work is it turns out that there can be worlds in which physicalism is true and yet there is an infinite descent of mental levels. How can this be?

1.3. A compositional account of mentality
My goal here is to describe a possible world with an infinite descent of mental levels, yet in which physicalism is true—to make this intelligible I will elaborate on what kind of mentality exists at every level of that world. Begin by considering a real-world brain: if physicalism is true, and eliminativism about the mental is false, in virtue of what does the brain get its mentality from? That is, why does this particular physical system—the brain—have mental properties?

It turns out that there is a commonality most physicalism-friendly theories of mentality share—these views will have the brain count as mental in virtue of the relational properties of the parts of the brain. “Relational properties” has a wide scope, and is intended to cover both stereotypically relational properties, like spatiotemporal and structural properties, as well as causal and dispositional properties, which are properties that govern the behavior of and interactions between things. I will call mentality of this sort “compositional mentality”.
Compositional mentality is a background condition for all physicalist theories of mind which are substrate-neutral—that is, all views which say mental properties depend on properties that are not specific to the lower-level realizers of those properties. This includes all varieties of role functionalism, as well as informational and representational theories, and—as I will show—may even include certain versions of identity theory.

Allow me to demonstrate with role functionalism. Role functionalism identifies mental properties with functional roles, where a functional role is a higher-order property that functionally describes some other property or properties at a level (see Witmer 2003)—the functional role of being the bait in a mousetrap is a property of the cheese, but is not identical to the property of being cheese (since, after all, it could have been played by peanut butter)—and which is defined by input-output relations and state transitions. For instance, the functional role of pain might be defined as anything that is caused by stimuli that we normally label as pain-causing, like bodily harm to enervated tissue, and which has output to appropriate other mental states—like the thought ‘damn that hurts!’—and to pain-avoiding behavior. This is ‘role’ functionalism as opposed to ‘realizer’ functionalism because it assigns mentality to the functional role rather than any particular lower-order realizer.

In the case of the brain, functional roles are realized by parts that are interacting in the right way. Once you get excited neurons spitting chemicals out at each other and creating networks of electrochemical neural activity correctly, you get functionally-characterized mentality for free. Of course, to get the neurons doing this you must arrange their parts—biochemical and atomic components—properly, and so too must arrange all of the lower level components, like quarks and electrons, in a roughly correct way as well. I say “roughly” because presumably below some level there can be changes in component arrangement without changes in mental state—these changes are in the so-called noise and are irrelevant to the higher-level functional role. This general account will be true for any mental thing, not just a brain: to realize a functionally-characterized mental state or property, just put the interacting parts—be they neurons, microchips or alien goo—together in the right way.

---

7 This is obviously a simplification of which and how many input-output relations are sufficient for realization of pain on any functionalist story, but the more elaborate account—perhaps involving Ramsey sentences that quantify over sets of named functionally-characterized mental properties for a given theory—will be more-or-less along the same lines.
One might object that this characterization of the world—involving breakdown of macro-level objects into discrete interacting components—sounds too naïvely mechanistic. It might eventually be revealed that, instead of discrete particle-like things, the world is composed of fields without well-defined boundaries, and that the low-level entities are more like clouds than billiard balls. And certainly if we look at the brain as a biological system we don’t see something like Leibnitz’s mill in miniature but, as Peter Godfrey-Smith (forthcoming) calls it, a “molecular storm” of activity. I certainly don’t want to commit, either at the microphysical or the biological level, to an overly simplistic and mechanistic view of nature—but luckily, I don’t think I am so committed.

The core features of the view here—that relational properties of constituents are what count for generation of physicalism-friendly mentality—can survive even if the components are field-like and boundary-less, or even weirder than that. Functionally-characterized mentality will still depend on interaction among parts, regardless of how those parts are described—and I think the brain case makes this eminently clear. For the brain, relational properties of the parts are doing the work to implement a functional state, regardless of what squishy, chaotic, messy kind of thing a neuron is.

Role functionalism relies on the relations among the parts of a physical system to generate mentality, and nothing further is required. The same is true for informational theories, which say that something is mental if it realizes appropriate informational properties, e.g. some minimum level of informational integration (Tononi 2007). Informational properties are substrate-neutral, and depend entirely upon structural and dynamical features of the system. So too for representational theories, though the story is a bit trickier: for most contemporary physicalists, some system is representational if its inner states reliably match states of the world, and some addition requirement is met, like the system has the function to perform representation due to its evolutionary history (e.g. Millikan 1989). Such representational properties are structural, though they depend on relations between structures in the mind and structures in the world, and sometimes enormously temporally or spatially extended structures (as under evolutionary versions of teleosemantics). Nonetheless, it is relational properties (again, broadly construed) which are relevant to the generation of mentality for these theories.

The odd theory out is identity theory, which in its most common form identifies mental properties with neural properties. On first pass, neural properties are not substrate neutral, and thus do not depend
for realization only on relational properties of the parts of the brain. It seems that even if the very same relational properties realized by the parts of a brain could be realized by a virtual brain in a digital simulation, that digital brain would not truly realize neural properties. Though this sounds right, there is a way to treat neural properties which would make identity theory consistent with compositional mentality: if all properties are individuated by their causal dispositions, as perhaps under dispositional essentialism, then neural properties are identical to complex combinations of lower-level relational properties. This allows neural properties to count as instances of compositional mental properties. However, this line of thinking is metaphysically speculative, and I’ll leave identity theory aside in what follows.

Though compositional mentality is a background condition for many theories of mind that are widely held to be consistent with physicalism, this view is also consistent with an account of mentality that is incompatible with physicalism: necessitarian dualism (White 2018, Brown forthcoming). Like typical physicalism, this nonstandard nonphysicalist view holds that mental properties are necessitated by relational properties of the parts of a system, but additionally asserts that the mental properties which are necessitated are non-physical.

I would like to use compositional mentality to generate an infinite descent of prima facie physicalism-compatible mentality. To avoid issues with necessitarian dualism, I will bracket this sort of mentality: the worlds which I will describe are worlds in which necessitarian dualism is false. Since compositional mentality is consistent with many other accounts of mentality which are widely recognized as physically kosher, an instance of (non-necessitarian dualist) compositional mentality in a world ought to prima facie count as mentality which is consistent with physicalism. To put it another way, once necessitarian dualism is ruled out, it would be quite surprising to find an instance of compositional mentality which renders physicalism false.

1.4. Quasi-brains all the way down
Now, equipped with this composition-dependent, physicalism-friendly notion of mentality, I can finally describe my all-mental and all-physical world—MPW for short. MPW is, of course, one with an infinitely descending series of levels—there are no smallest entities nor basic properties that this world is ultimately grounded in. Moreover, every level has a high compositional complexity—thinking of levels mereologically, any entity at any level n has many constituent parts at the next lowest level, n-1. How
many? At least as many as are required to realize minimal mentality for the thing that the parts compose. If mentality comes cheaply, as some generous functionalists have suggested, then entities might not need to decompose into terribly many parts; but if mentality requires a high degree of complexity, then each entity is composed of however many parts are required for such complexity.

Next imagine that MPW contains a single highest-level system—I say system to indicate that it is something with a complex mereological structure and dynamic causal relationships among its parts—called Alex. Nothing else exists in this world apart from the single, internally sophisticated thing under consideration. Further, upon inspection of the system and its properties we find—lo and behold!—it is a thinking (or feeling/experiencing) thing. That is, Alex has mental properties.

Why is Alex mental? For the same kind of reason you are: Alex has parts that are dynamically arranged in such a way that they produce mentality for the thing they compose. As I said above, I believe that this is the common feature of nearly all physicalism-friendly stories about mentality. For instance, if functionalism is true, then Alex gets mentality because the constituents of Alex at the level immediately below are causally related just so that they realize functional roles.

This gives mentality to the thing at the highest level of the world, but what about my promise of making mentality ubiquitous, all the way down? Well, just iterate the story I told about Alex for all of Alex’s parts. Alex is made of many interacting components, call them Bretts. Collectively all the Bretts, through their interactions, generate Alex’s mentality. But also each Brett is mental, for the same reason Alex is. Remember that every entity has high compositional complexity—every Brett has many parts, call them Charlies, which cause the Brett to be mental in virtue of the Charlies’ interactions. And the same story goes for each Charlie which is composed of little Devins, and so on infinitely.

It is sort of like if you could replace your brain’s neurons with tiny brains that are connected to each other and which realize the same functional relations your neurons had with each other, and then also replace the neurons of the constituent neuron-replacing brains with more, littler brains, and so on and on and on (see Block 1978 for a similar case). A (Alex) is mental in virtue of the relations between the Bs (Bretts) that compose A, and the Bs in virtue of the relations between the Cs, and Cs the Ds, without end for all of the infinite pairs of levels. The most notable difference between role functionalist and type-identity accounts of mentality in this system is that some particular mental property can be realized at
different levels—for Alex, Bretts, Charlies and so on—under role functionalism, and mental properties are level-specific under type-identity theory, such that Alex and a Brett can never have exactly the same mental property.

There it is—a world that is completely mental at all of its infinite levels of decomposition. Every entity at every level has mental properties in virtue of the next-lower-level compositional structure of that entity. But why should you agree that physicalism could be true in MPW? After all, it violates one of the criteria given above for when a world is to count as physical: no infinitely descending all-mental levels. Of course this is exactly the principle that I am challenging, so an argument is required to show that it does not count as physical. Yet, I see none in the offing.\(^8\)

In MPW mentality is generated at every level through the dynamic compositional structure of lower level entities—remember that this is how all physicalism-friendly accounts of mentality work. In essence, I am just positing many, many quasi-brains—brains all the way down, if you will. The substantive issue cannot be the sheer volume of mentality in MPW, because we can imagine an infinitely large universe nomically similar to ours, but with an infinite number of humans in it: such a world has an infinite number of mental properties, yet it is conceivably physical. And to say that MPW is nonphysical only because it contains an infinite descent of mentality is clearly question-begging.

Alternately, one might argue that physicalism is true in MPW, and that what I have described is not genuinely in tension with the no-infinite-descents-of-mentality criterion. There are various ways one might do this. One might think that physicalism entails eliminativism—not a very popular route for a physicalist, but it would trivially mean that no way of arranging only physical things could produce mentality. Given that most physicalists are not eliminativists, and that MPW straightforwardly contains physicalism-friendly mentality, the other option is to say that the kind of mentality in my world is not the

---

\(^8\) Additionally, one might wonder about the relationship between MPW and panpsychism: panpsychism is typically understood as the thesis that all fundamental entities have both physical and mental properties. This conception is problematic because, like physicalism defined in terms of fundamentality, it has panpsychism come out as trivially false in worlds with an infinite descent of levels. It seems that any plausible revision to panpsychism which allows it to be true in a world without a fundamental level will have panpsychism come out true MPW: if panpsychism is revised into the thesis that mental and physical properties are ubiquitous at every level of nature below some level, then panpsychism so conceived is true in MPW. However, it seems that panpsychism might be false in MPW—after all, the mental properties in MPW are characterized as consistent with physicalism. This is the inverse of the problem that MPW raises for physicalism—panpsychism is true in MPW when it should come out false, versus physicalism being false in MPW when it should come out as true—and to me it seems just as difficult to resolve.
kind of mentality that the physicalist had in mind when she came up with her sufficient-for-false condition. So a slight amendment to the criterion is in order, to make it say something like ‘only an infinite descent of a certain kind of mentality is sufficient to make physicalism false’.

How should this physicalism-hostile kind of mentality be characterized? One might be tempted to say that the so-called “bad mentality” is any kind other than that characterized by a particular physicalism-friendly theory of the mental. A strategy like this is pursued by Tiehen (2015), who says that physicalism is committed to functionalism: if he is right, then only functionalist mentality can exist in a physical world, and no amount of this sort of mentality can make physicalism false—even an infinite descent of it. However, I have already shown that this response won’t work for the world I have described. Mentality in MPW is consistent with various theories of mentality—it contains an infinite descent of functionally-characterized mentality, as well as informationally-characterized mentality, and perhaps representationally-characterized mentality. So exempting mentality as described by a particular acceptable theory won’t make physicalism true in my world, since MPW also contains an infinite descent of differently-characterized mentality.

But maybe there is a way to characterize categorize types of mentality more generally, such that an infinite descent of any physicalism-friendly kind of mentality is not sufficient to make physicalism false. In fact, it seems I have provided just such a characterization—perhaps an infinite descent of mental properties that depend for realization only on structural-dynamic arrangement of constituent parts is consistent with the truth of physicalism.

That seems promising, but it won’t work either—there are anti-physicalist accounts that say mentality depends only on structural-dynamic arrangement of constituents. Emergent dualism says that arranging low-level things such that they have appropriate structural-dynamical relations to each other generates emergence of non-physical mental properties. If physicalism ought to be false in such a world, then the constitutive account of mentality is perhaps a necessary component of physicalism-friendly mentality, but not sufficient.

A different critique is to say that nothing in MPW is genuinely mental. If so, then MPW doesn’t contain an infinite descent of mentality, and there is no issue for physicalism. One might think this because she believes that no mental thing can have another mental thing as a part, perhaps on the
grounds that such systems as Block’s (1978) Chinese nation or Searle’s (1980) Chinese room shouldn’t count as genuinely mental even if functionalism is true. Both of these fictional systems are mereologically and functionally sophisticated, enough to be mental on most functionalist accounts, but they have human beings as constituents that are involved in functional processes.

One response to this objection comes from cognitive science: arguably the human brain is composed of mental sub-systems. This is doubly evident if any modularity thesis is true—which would mean, as in Fodor (1983), that there are process specific, hard-wired, informationally encapsulated and cognitively impenetrable mental sub-systems that compose more general human mentality (in particular in the perceptual system). If this is right, then it seems a physicalist who denies that mental systems can have parts that are mental must deny mentality to the human brain—surely not an acceptable consequence. Perhaps, one might respond, these sub-personal systems should be deemed not truly mental, but then what is the criterion a physicalist can give for when something is to count as mental? This is especially worrying if these systems are treated as genuinely mental by cognitive scientists.

A different kind of response comes from Eric Schwitzgebel (2015), who argues that there are nomically possible alien creatures who could be constructed out of undeniably mental parts, and who would be mental. He argues that it would be chauvinistic for humans to deny that such beings—some of which might resemble incredibly sophisticated ant colonies—could have mentality, given they would exhibit all the other hallmarks of having mental lives, for instance the ability to have conversations about what kind of tea they most enjoy the taste of. Further, such beings must have mental properties if physicalism or materialism is true, since all physicalism-friendly accounts of mentality say that there is nothing more to mentality than the properties realized by these odd systems—and this is consistent with the treatment of mentality I am offering.

As an additional consideration, one might wonder about the relationship between MPW and panpsychism: panpsychism is typically understood as the thesis that all fundamental entities have both physical and mental properties. This conception is problematic because, like physicalism defined in terms of fundamentality, it has panpsychism come out as trivially false in worlds with an infinite descent of levels. It seems that any plausible revision to panpsychism which allows it to be true in a world without a fundamental level will have panpsychism come out true in MPW: if panpsychism is revised into the thesis
that mental and physical properties are ubiquitous at every level of nature below some level, then panpsychism so conceived is true in MPW. However, it seems that panpsychism might be false in MPW—after all, the mental properties in MPW are characterized as consistent with physicalism. This is the inverse of the problem that MPW raises for physicalism—panpsychism is true in MPW when it should come out false, versus physicalism being false in MPW when it should come out as true.

One way to resolve this tension is to give up the idea that panpsychism and physicalism are inconsistent with one another; this raises an issue with Jessica Wilson’s (2006) response to Stoljar’s (2001b) argument that panpsychism is consistent with physicalism. Stoljar claims that panpsychism merely posits more and stranger mental properties than we typically suppose exist, but that quantity and location of mentality alone is not enough to make physicalism false. Against this, Wilson argues that physicalism must contain a “No Fundamental Mentality” (NFM) constraint—this is the same as the via negativa account of physicalism, but employed as a necessary condition rather than a definition, and she thinks this principle explains why panpsychism is inconsistent with physicalism. However, the sibling to Wilson’s NFM constraint which would apply to worlds without an infinite descent of levels is Montero’s no infinite descent of mental properties constraint—yet, as I have been arguing, there are plausibly physical worlds that violate such a constraint. Perhaps this indicates, contrary to Wilson, that some notion of panpsychism is consistent with physicalism.

1.5. Conclusion

The argument presented here is aimed at via negativa physicalism, which says that physicalism is false in a world if that world realizes fundamental or an infinite descent of mental properties—and so far this is the only dependence-based account which allows physicalism to be true in infinitely decomposable worlds. The problem, which I exploited to build an infinitely descending all-mental and all-physical world, is that ‘mental’ is underspecified. In effect, via negativa physicalists are defining an opaque term with an even more opaque term—there is even less agreement over what ‘mental’ means than what ‘physical’ means. However, via negativa’s cloud has a silver lining, since it seems this problem might be resolved by being more careful about what is meant by ‘mental’. Doing so might supply an unexpected boon to via negativa—there is something oddly insubstantial about via negativa’s definition of physicalism; after all, it does not positively define what the physical is, but only what it is not. Fleshing out a general physically-
acceptable notion of mentality would answer not only the problem posed in this dissertation, but perhaps would add the meat that is lacking from via negativa’s bones.

2.1 Fundamental mentality in physical worlds

Nearly everyone accepts that the existence of fundamental mentality would render physicalism false. I believe this widely held view is false: there is a class of possible worlds which contain fundamental mentality and in which physicalism is plausibly true. These are worlds in which priority monism (Schaffer 2010)—which is the view that the properties of the whole of nature are fundamental, rather than the properties of the smallest parts—is true, and in which the whole of nature is a mental system of a certain sort. Schaffer (2010, p. 31-32) describes a priority monist as saying that:

[…] the whole is prior to its parts, and thus views the cosmos as fundamental, with metaphysical explanation dangling downward from the One […] The core tenet of historical monism is not that the whole has no parts, but rather that the whole is prior to its parts. […] The historical debate is not a debate over which objects exist, but rather a debate over which objects are fundamental. I will defend the monistic view, so interpreted: the world has parts, but the parts are dependent fragments of an integrated whole.

Priority monism is distinguished from what I will call “priority partism”, which says that the smallest parts of the world and their properties are fundamental. To illustrate: if everything is ultimately composed of subatomic particles, then priority partism deems that subatomic particles and their properties are fundamental. Priority monism, in contrast, would assert that the properties of subatomic entities ultimately depend on the properties of the whole cosmos.

Since any properties of the whole of nature trivially count as fundamental under priority monism, it follows that the mentality of a world-encompassing mental system is trivially fundamental in a priority monist world. If this mentality is of a physically acceptable sort, then physicalism ought to count as true in some of these worlds—despite the existence of fundamental mentality! For instance, in a priority monist world containing only a single functioning brain—and absolutely nothing else—the mentality of that brain
trivially counts as fundamental. Yet, at a glance, its mentality seems physically unproblematic. I will describe several such worlds in a way that will draw out the plausibility that physicalism is true in them.

As I see it, the problem with via negativa and No Fundamental Mentality is that they are too crude: beyond a fundamental/non-fundamental contrast, these theses do not sufficiently make a distinction between mentality that is physically acceptable versus mentality that is physically unacceptable. As I argued in the first section, this leads to problems in certain worlds which are infinitely decomposable, such that they contain no smallest parts and—assuming priority partism—no fundamental properties. There is a class of such worlds which contain an infinite descent of mentality, yet in which physicalism is plausibly true, in virtue of the sort of mentality that those worlds contain. Attention to priority monism reveals a similar result in worlds containing fundamental mentality.

Here is the roadmap for what follows: section 2.2. will describe worlds in which priority monism is true and which realize compositional mentality at the level of the whole of nature. Section 2.3. will respond to anticipated objections, and section 2.4. will conclude.

2.2. Universe-sized Minds in Priority Monist Worlds

Priority monism is an unconventional metaphysical thesis about what properties in a world are fundamental. This view has received its best contemporary articulation and defense by Jonathan Schaffer (2010), in which he defines priority monism as the view that the properties of the whole of nature are fundamental, with the properties of the parts of the universe ultimately dependent on the properties of the whole. This is in contrast with the traditional view—priority partism—which deems that only the properties of the smallest things in existence are fundamental. The traditional view received an early explicit articulation by Putnam and Oppenheim (1958), and has been taken for granted by analytic metaphysicians until only recently.

Schaffer contends that if priority monism is true in any world, then it is true in all worlds. He additionally provides several examples and arguments involving massive quantum entanglement, infinite decomposability and a well-groundedness principle (which is the idea that all chains of dependence must ultimately be grounded in fundamental properties) to demonstrate that priority monism is true in some
The first, largely assumed contention of Schaffer's is unnecessary for my argument. I take it not everyone will agree that if priority monism (or some other metaphysics of dependence, for that matter) is true in any particular world, then it must be true in all worlds (see Siegel 2016). Regardless, all that matters here is that priority monism could be true in some world.

Why think priority monist worlds are possible? Here is a reason: priority monism does not seem to be a contradictory thesis—after all, there does not seem to be anything in the concept of fundamentality to deem that only the properties of the smallest things in existence can be fundamental. If this is right, then a world in which priority monism is true is, at minimum, negatively conceivable—being conceptually non-contradictory is sufficient for this sort of conceivable (I'm borrowing from the terminology of Chalmers 2010). If, as Chalmers argues, negative conceivable entails possibility, then priority monism is metaphysically possible. Of course, not everyone agrees that the non-contradictoriness of a conceived state of affairs entails that such a state of affairs is possible. So, I must admit that my overall argument rests on the conditional 'if priority monism could be true in some possible world, then it is possible for there to be physically acceptable fundamental mentality’—but the antecedent is not flagrantly suspect, in my eyes.

On the assumption that priority monism is metaphysically possible, let's look at the details of how compositional mentality—which I described in section 1.3.—could be realized at the level of the whole of nature in a priority monist world. This is fairly easy. Imagine a world containing only a single functioning brain. All non-brain entities in this world are parts of the brain—nothing is left floating free in the cosmic void. This brain realizes mental properties, and those mental properties are necessitated by the realization of appropriate relational properties of the parts of the brain. If priority partism were true in this world, then the mental properties of the brain would not count as fundamental. Instead, the properties of the most basic constituents of the brain, e.g. the charge and spin of its electrons, would be fundamental.

The gist of these arguments follows. Priority partism cannot be true in worlds which are infinitely decomposable, since these worlds contain no smallest entities which would count as fundamental. The well-groundedness principle says that all chains of dependence must terminate in fundamental properties, and the only way to do this in infinitely decomposable worlds is if priority monism is true in them. I discuss quantum entanglement at greater length in the next section, but the idea of this argument is that the properties of quantum entangled systems supervene on the properties of the whole system, and not vice versa. Any worlds which are “massively quantum entangled”, such that everything is entangled with everything else, are worlds which would have the properties of all the parts ultimately dependent on the properties of the whole, and thus the whole should count as fundamental in these worlds.
However, add the detail that priority monism is true in this world and—presto! The mental properties of the brain—which is the sole inhabitant of the world, so its properties are the properties of the whole of nature—automatically count as fundamental.

As I said, easy. There seems to be nothing physically untoward about the mentality that is realized in this world: the brain I have described is nothing like a Cartesian soul in any sense. Panpsychism is not true in this world, nor is the priority monist version of panpsychism called “cosmopsychism”, as has been developed by Philip Goff (2017). Aside from not being connected to a body, the brain in this world is not in any significant respects dissimilar from an ordinary brain in the actual world. The only things that make a metaphysical difference between the mentality in this imagined world and ordinary mentality in the actual world are: priority monism is true in the imagined world, and the imagined world only contains a solitary brain.

Allow me to provide one more example to flesh out this class of worlds, which will additionally demonstrate some strange consequences that follow from the No Fundamental Mentality constraint on physicalism. Imagine a world in which priority monism is true, which is finite in size, and which contains a species of intelligent beings like humans. The intelligent beings in this world create a “hungry AI” which has one goal: to incorporate everything that exists into its computer hardware, in order to increase its cognitive powers. The AI is extremely successful, first turning organisms and rivers and mountains into parts of its computer hardware, followed by planets and stars and galaxies. It does this by reducing objects to their constituents, then using those constituents to expand its hardware resources.\(^\text{10}\)

Eventually, the AI transforms nearly everything into parts of its cognition-realizing computer hardware. The only thing which remains outside the AI system is a stray electron that has not yet been incorporated. Prior to incorporating this lonely electron, the whole of nature does not realize mental properties—rather, only a large part of nature does.

Recall that priority monism is true in this world. When the AI finally incorporates the last electron, and the whole of nature becomes an AI system, then suddenly the mentality of the AI counts as fundamental. This is because under priority monism, the properties of the whole are fundamental by definition, regardless of whether those properties are mental or not. According to via negativa or No

\(^\text{10}\) This is a scenario that some AI ethicists worry about in the actual world, e.g. Bostrom (2014) and Schneider (2018).
Fundamental Mentality, physicalism will have gone from true to false in this world. Yet this is not a world containing souls or panpsychic mind dust. It is just a universe containing a very powerful and large computer system. It would be absurd to maintain that physicalism has become false in this world.

The worlds just described realize mentality that is of the compositional variety, and necessitarian dualism is stipulated as false in them. They additionally contain fundamental mental properties—thus making physicalism false if one accepts a No Fundamental Mentality constraint on physicalism. Yet physicalism plausibly ought to count as true in these worlds, as well as an infinite number of others in their class.

2.3. Objections

There are a number of objections available to those who think that the existence of fundamental mentality is inconsistent with physicalism. However, short of fleshing out a more general account of physically acceptable mentality that does not rely on a simple fundamental/nonfundamental distinction, I don’t think any of these objections stick.

First: one might object that priority monism is impossible. This might be on the grounds that priority monist worlds are ultimately inconceivable (perhaps there is a subtle contradiction in the view) or on some other grounds. As I already admitted, this response would successfully undermine my argument: if there are no possible worlds in which priority monism is true, then there are no possible worlds containing physically acceptable fundamental mentality such as I have described.

I don’t have much to say in response to this objection beyond reasserting that priority monism seems possible. I admit this might be wrong—after all, priority monism is certainly quite weird, and many of us are naturally inclined toward priority partism (though perhaps the inclination is due to philosophical training). However, it seems to me that the onus is on an opponent of priority monism to provide a strong argument that it is impossible. And as yet no such argument has been articulated.

Alternately, one might think that priority monism is possible, but only if some further condition is met—specifically, priority monism is only true in worlds which are massively quantum entangled.\footnote{Ishmael and Schaffer (2016) further develop the position that worlds which are massively quantum entangled are worlds in which priority monism is true, arguing that the parts of entangled systems are nonseparable, which means that their states cannot be specified without referring to each other. They think that this forces us to accept that the properties of the parts of such nonseparable systems depend on the properties of the whole system. Calosi (2018) argues that only certain}
argument is undercut if this additional condition makes physicalism false, or if it is inconsistent with realization of the sort of compositional mentality I described in section one. If massive quantum entanglement is inconsistent with physicalism, then there is no sense in worrying about physically acceptable fundamental mentality in those worlds, since physicalism would be false in them regardless. Alternately, if massive quantum entanglement is inconsistent with whole-world compositional mentality, then worlds such as those I described in the last section turn out to be impossible.

The basic idea motivating this objection is plausible: something seems required to make priority monism rather than priority partism (or some other metaphysics of dependence) true in a world. It would be bizarre if priority monism could be true in some world, and priority partism true in some other, with no other difference between them.\(^{12}\) Even with this granted, neither of the subsequent claims of the objection look right to me: no such condition by itself seems sufficient either to make physicalism false or preclude the possibility of compositional mentality being realized for the whole of nature.

Suppose that priority monism is only true in worlds which are massively quantum entangled. In such a world, it is nomologically impossible for some part of the world to change without all the parts of the world also changing—but this is just a consequence of quantum entanglement, not its essence. Rather, Schaffer (2014) defines quantum entanglement in the following way:

An entangled system is one whose state vector is not factorizable into tensor products of the state vectors of its components:

\[
\Psi_{\text{system}} \neq \Psi_{\text{component1}} \otimes \Psi_{\text{component2}} \otimes \Psi_{\text{component3}} \otimes \ldots
\]

What this inequality means is that the quantum state of an entangled system contains information over and above that of the quantum states of the components. The intrinsic properties of entangled wholes do not supervene on the intrinsic properties of and spatiotemporal relations among their parts.

\(^{12}\) Note that grounding theorists who posit metaphysical grounding laws (e.g. Wilsch 2015) need not require this: there could be two worlds indiscernible from one another save that one has priority monist grounding laws, and the other has priority partist grounding laws. Regardless of the plausibility of this sort of metaphysics of grounding, my response to the objection does not rest on the existence of occult grounding laws.
In other words, if something is part of a quantum entangled system, then at least some of its properties are necessitated by the properties of the whole system, and the properties of the whole system are not necessitated by the properties of the parts of the system. Is a massively quantum entangled world inconsistent with either physicalism or whole-world compositional mentality?

I believe that the answer is “no”, but this is a complicated issue. One might think that a massively quantum entangled world is physically unacceptable because it entails the existence of strong emergent properties. Strongly emergent properties are not necessitated by, nor a priori inferable from, the intrinsic and spatiotemporal properties of a thing’s parts. Though some have argued that strong emergence of this sort is incompatible with physicalism (Jackson 1998), it is not clear why it should be. After all, our best physics says that the actual world contains quantum entangled systems (and it even might be massively entangled, as Schaffer 2010 argues), and it is strange to think that physicalism could be made false just for this reason. Because of this, it would be a misstep to assert that physicalism is inconsistent with the mere existence of strongly emergent entangled properties (see also Papineau 2008, Brown and Montero 2018).

Whether quantum entanglement is consistent with compositional mentality is a thornier question. Remember that compositional mentality is realized if relational properties of the parts of a system necessitate mental properties for that system. If the properties of a whole quantum entangled system fail to supervene on the properties of the parts of a quantum entangled system, then it seems that the properties of a whole quantum entangled system cannot be identical to compositional mental properties. Given how quantum entanglement is defined, this worry seems reasonable. However, rather than get into the nitty gritty of the metaphysics of quantum mechanics, I would like to undercut this objection in a more straightforward way.

An actual human brain is quantum entangled system, but it nonetheless realizes compositional mentality. This is because the mentality-relevant properties of a brain—like biological and chemical properties of neurons—are not affected by quantum effects. As many have pointed out (e.g. Tegmark 2000, Jumper and Scholes 2014), the brain is too hot and wet for intrinsic quantum effects to make a difference to its mentality-relevant operations. So, even though the brain is quantum entangled, its mental properties are necessitated by the relational properties of its parts. Since the world I asked you to imagine
in the last section contains a brain that is just like an actual-world brain, it too should be capable of realizing compositional mentality, regardless of whether it is a quantum entangled system or not.

However, I must stress that this is a complex issue. There may be a flaw in my response—for instance, perhaps it will turn out that human brains regularly employ quantum mechanical effects in their operations (see Penrose 1989, 1994; Hameroff and Penrose 1996). Or perhaps there is some deeper metaphysical problem that I am overlooking. Regardless, even if I am wrong on this, the overall point I am endeavouring to prove remains correct. Consider the hungry AI world that I described in the last section. Suppose that in a massively quantum entangled world, such an AI is incapable of realizing compositional mentality—instead, its mentality is a strongly emergent property that is not necessitated by the relational properties of the parts of the AI system (perhaps it is a quantum computer).

Possession of this sort of mentality by the AI is surely consistent with the truth of physicalism. Call mentality that is identical to the emergent properties of a whole quantum entangled system “entangled mentality”. If entangled mental properties depend on quantum mechanical laws which are not especially suited to produce mentality, then those properties are not physically pernicious. And entangled mental properties do depend on quantum entanglement laws which are not mentality-specific—these laws work equally well to generate entangled non-mental systems, like tables and chairs. Compare entanglement laws with the physically problematic psychophysical laws of Chalmers (1996), which exist just to nomologically generate mental properties given appropriate physical conditions, and are incapable of generating non-mental properties. Additionally, since brains are quantum entangled systems, it would again be foolish for a physicalist to deem that the existence of entangled mentality should make physicalism false.

Moving away from issues with quantum entanglement, an objector might take a different tack, and assert that there are two sorts of fundamentality: priority partist fundamentality and priority monist fundamentality. The No Fundamental Mentality thesis concerns only the former sort, so the possibility of priority monist fundamental mentality does nothing to harm the thesis.

This sounds wrong to me. There seems to be only one sort of absolute (as opposed to relative) fundamentality: some property is fundamental if it depends on nothing else. This is necessary and
sufficient for a property to count as fundamental, regardless of whether it a property of the smallest part or the whole of nature.

However, for the sake of argument, suppose there really are two ultimately different kinds of fundamentality. Does this meant that No Fundamental Mentality is off the hook? I think not: consider Goff’s (2017) cosmopsychism. This is the view that the whole of nature possesses fundamental mentality, but not because compositional mentality (or some other physically acceptable sort of mentality) is realized. Rather, Goff’s view is somewhat similar to panpsychism, which says that the most mereologically basic constituents of reality possess a primitive sort of mentality. Under consmopsychism, the whole cosmos possesses a primitive and scientifically inexplicable sort of mentality, with the mentality of the parts of the universe depending on the mentality of the whole.13

To rule out a view such as cosmopsychism, a “No Priority Monist Fundamental Mentality” constraint must be posited in addition to the standard “No Priority Partist Fundamental Mentality” constraint. Unfortunately, the new constraint makes the wrong ruling on worlds in the class which I have been describing. So stipulating that priority monist fundamentality is distinct from priority partist fundamentality does not help.

One might instead bite the bullet, and admit that physicalism is false in worlds such as those I have described. After all, priority monism is weird, so why trust our intuitions about the conditions under which physicalism is true in those worlds? It just turns out that our common sense leads us astray in these cases.

This looks to me like the worst response: the bullet is quite big, with unexpected consequences that I have already identified, such as a hungry AI making physicalism go from true to false in a world. It is also clearly question begging: this objection just denies the conclusion of my argument without wrestling with any of the premises. The objection is particularly bad because there seems to be nothing physically unacceptable in the worlds I have described: priority monism is consistent with physicalism, and

---

13 Goff attaches cosmopsychism to a version of panpsychism, such that all of the components of the whole also possess mentality. I don’t see why cosmopsychism and panpsychism must be linked, but regardless, a different problem case can be contrived which combines the infinite descent of mentality from the last section with the priority monist mentality of this section. This is a world which contains both cosmos-level mentality (which is fundamental under priority monism) and mentality for all the parts of the cosmos. Thus, if one were to object that the priority monist mentality I have described is different than the priority monist mentality Goff describes, a truer analogue of Goff’s cosmopsychism can show up in the world which combines priority monism with an infinite descent.
compositional mentality is consistent with physicalism, and it really looks as if the conjunction of priority monism with compositional mentality ought to count as consistent with physicalism. The objection could be given teeth if a good reason were given for why physicalism ought to count as false in these worlds. However, I see no such good reason.

Finally, perhaps the No Fundamental Mentality proponent might admit that physicalism is true in the worlds I have described, and true precisely because compositional mentality is realized in them. This objector might urge that I have provided the solution to my own puzzle: what is needed is a more specific account of physically acceptable versus physically unacceptable mentality, and compositional mentality provides just such a condition. After all, what other sort of mentality could be consistent with physicalism?

Unfortunately, this doesn’t work: as I have already noted, compositional mentality is neither necessary nor sufficient as an account of physically acceptable mentality. Russellian physicalist mentality (Montero 2010, Brown 2017)—which says that phenomenally mental properties ultimately depend on non-relational properties of the basic constituents of the world—seems to be a physically acceptable sort of mentality. Yet it is inconsistent with compositional mentality: appropriate relational properties of the parts of a system are not sufficient to necessitate Russellian physicalist mentality. If this is right, then compositional mentality is not necessary in order to have physically acceptable mentality. Similarly, as I explained already, necessitarian dualism is consistent with compositional mentality. So compositional mentality is not sufficient for physically acceptable mentality either.

2.4. Conclusion

Though ‘physicalism’ is underdefined, nearly all philosophers agree that the existence of fundamental mentality would make the view false, regardless of whatever else physicalism requires. I have presented a challenge to this widely accepted attitude: fundamental mental properties could exist which should not make physicalism false. What does this mean for our definition of physicalism?

I agree with those philosophers who think that physicalism should be made false by instantiation of physically unacceptable mental properties. It is a social fact that most philosophers understand certain views about e.g. consciousness to be inconsistent with physicalism. The problem is that characterizing physically unacceptable mentality as mentality which is fundamental does not suffice—such an articulation forces us to make the wrong ruling on a class of worlds that plausibly ought to count as
physical. Some more sophisticated articulation of physically acceptable mentality needs to be given if anything like the *via negativa* account, or a *via negativa*-plus-something-else account, is to work.

As yet, no such articulation has been offered. This is not a simple task: such an account of physically acceptable mentality must rule out views like necessitarian dualism, and include views like Russellian physicalism. I do not yet have a positive view to offer as aid for via negativa physicalists. Nonetheless, I expect that there is a way to articulate the difference between physically acceptable and physically unacceptable mentality in a way that does not simply appeal to fundamentality, and that this can be used by *via negativa* physicalists as an improved version of their view. Now the hard work has to be done to articulate such an account of the mental.
II. The Hope and Horror of Physicalism

Philosophers’ passions are frequently aroused by arguments physicalism. Though physicalism stands as orthodoxy in analytic philosophy, claiming to be the metaphysical position that best suits our scientific age, one might wonder why anyone other than an analytic philosopher should care about the thesis. After all, the undeniable success of science is independent from the truth of physicalism, and nothing of obvious consequence to non-philosophers appears to stand or fall with the view.

In contrast with this deflationary attitude, physicalism can be considered as an existential condition, which is a state of affairs that is relevantly related to the existential question ‘should I live given condition x?’ The relevance relation is one of commonsense intelligibility: if it makes sense to ask the existential question for some particular condition x, then that condition is existential. ‘Should I live given that yellow is not my favorite color?’ would probably not count as plausible by anyone’s standards, whereas ‘should I live given that everyone I love has died and I am suffering from late-stage Alzheimer's?’ would for just about everyone. Existential conditions come in two varieties: positive conditions, which prima facie lend themselves to an affirmative answer to the existential question, and negative conditions, which lend themselves to a negative answer. For instance, the claim that ‘human life is brief and full of suffering in an absurd universe’ seems prima facie to lend itself to a negative answer, and it makes sense to question the value of continued personal existence given that condition. Alternately, ‘I am happy, healthy and wise’ lends itself to a positive answer.

There are several consequences of physicalism that appear to have existential import. On one hand, the view seems to be inconsistent with a number of things that many people find personally valuable, like the existence of normatively-laden laws of nature such as karma, the existence of a divine being such as God, the existence of an immortal soul, an ultimate teleological purpose to the universe or individual life, or the possibility of free will in a robust sense. On the other hand, physicalism also seems to have some positive existential implications, such as there being no limit to our potential scientific

---

14 This is an attempt at articulating the concern of Albert Camus when he says: “There is but one truly serious philosophical problem, and that is suicide. Judging whether life is or is not worth living amounts to answering the fundamental question of philosophy” (1957, pg. 1). I admit that there may be alternative ways to precisify the notion of an existential condition, but all such precisifications hopefully deem more-or-less the same things as existentially relevant.
understanding of the world, or support for the Epicurean attitude that we need not fear supernatural powers or forces of any sort (since they don’t exist). So at first pass, physicalism looks like it could be either a positive or a negative existential condition, depending on which of its consequences one focuses on.\textsuperscript{15}

However, whether physicalism really counts as an existential condition depends on the precise definition of ‘physicalism’, since the view is widely acknowledged to have a range of plausible interpretations. This determines which of the above existentially relevant consequences of physicalism really are consequences of physicalism. Are all versions of physicalism really inconsistent with the existence of robust free will, an immortal soul, God, ultimate purpose and human-centric laws of nature? Do they all support our scientific-imperialist hopes and allay our fears of supernatural forces?

I will focus on two relatively popular ways of understanding ‘physicalism’: the via negativa approach (Spurrett and Papineau 1999, Levine 2001, Montero 2005, Montero and Papineau 2005, Worley 2006) which understands ‘physical’ to mean ‘not fundamentally mental or fundamentally normative’; and the theory-based approach (Smart 1978; Melnych 1997, 2002), which says that something is physical if it is referred to by our best physics. Unfortunately, my task is complicated by flaws that have been identified with both approaches. As I discussed in the previous section of this dissertation, via negativa physicalism underspecifies the meaning of ‘mental’, such that it cannot distinguish between “physicalism friendly” sorts of mentality and “physicalism unfriendly” sorts, which leads to apparently wrong rulings on certain classes of worlds as being physical or not. The theory-based approach faces Hempel’s (1949, 1969) well-known dilemma: if ‘physics’ refers to current physics, then physicalism is false, since we know that current physics is not complete. If ‘physics’ refers to future or ideal physics, then physicalism is vague to the point of meaninglessness, or trivially true if ideal physics refers to any fundamental thing that could possibly exist (including paradigmatically non-physical things like God or Cartesian souls).

\textsuperscript{15} This list of the existentially relevant consequences of physicalism is different from what philosophers typically identify as the significant consequences of physicalism. For instance, Barry Loewer (2001) identifies fear of prioritizing physics over other sciences, reduction of everything to physics-level properties, issues about mental causation, and the threat of elimination of the mental as the prima facie scary consequences of physicalism. My list does not include any of these, since I think these are not issues most non-philosophers are gripped by. I also think Loewer is quite right that physicalism does not have these consequences—so even if people would be concerned (if properly informed about the matter) about e.g. the elimination of mentality under physicalism, the fear is ultimately unfounded.
I will argue that further precisification is required in order to assess whether the aforementioned existential consequences of physicalism are genuine consequences. I will offer some suggestions for the theory-based account, and provide a way to temporarily avoid the issues with *via negativa* physicalism. Specifically, while the *via negativa* might eventually relieve its woes by determining more precisely what counts as “physicalism friendly” mentality, here I will bracket off the classes of problematic worlds that the current formulation of *via negativa* makes the wrong ruling on. Theory-based physicalism might avoid the vagueness/trivial truth horn of Hempel’s dilemma by specifying the essence of scientific practice.

Finally, I will show that *via negativa* physicalism entails all (or nearly all) of the negative existential conditions identified above, and theory-based physicalism entails nearly none of them (but is nonetheless consistent with them). The positive existential conditions are split: theory-based physicalism satisfies our scientific hope to potentially conquer all of nature (in understanding at least), and *via negativa* largely abolishes the fear of supernatural forces as controllers of human destiny. Even though *via negativa* physicalism carries the weight of negative existential terror with it, this is not meant to be a knock against the view: this just means that the *via negativa* is a metaphysics that would have us confront reality without the artificial support of life-affirming myths.

I will proceed as follows. Section one irons out issues with the two popular ways to understand physicalism, and the following sections determine what the existential consequences of those two versions of physicalism are. I’ll conclude by offering some thoughts about what physicalism’s existential consequences mean for us and mean for theory selection in contemporary metaphysics.

1. Two Notions of the Physical

There is a hurdle to overcome in advance of applying the notion of an existential condition to physicalism: the two popular ways to define ‘physicalism’—theory-based physicalism and *via negativa*—each has its own set of definitional issues.\(^{16}\) Here I’ll explicate these issues in more detail, and offer some ways to patch each.

\(^{16}\) Additional ways to understand ‘physicalism’ have been proposed, for instance Ney’s (2008) view that physicalism is an attitude rather than a metaphysical thesis, or Stoljar’s (2001) “object physicalism” which says that all of the fundamental properties that exist are properties of prototypical physical objects. However, I take it that metaphysical accounts of physicalism are far more popular than Ney’s physicalism-as-attitude view, and in footnote ten of Stoljar’s (2001) *Two Conceptions of the Physical* paper he admits that object physicalism ultimately collapses to *via negativa* physicalism. Therefore, it seems reasonable to me to focus on theory-based and *via negativa* physicalism as the two standard interpretations.
As I already pointed out in the historical preliminaries and in the previous section, Carl Hempel saw that theory-based physicalism, which ties ‘physical’ to ‘physics’, is too ambiguous to be of any use: it gives us a view that is trivially false, hopelessly vague, or trivially true. One might think that via negativa physicalism is safe from ambiguity-related issues, but also in the last section I identified a definitional problem for via negativa: there are worlds that intuitively ought to count as entirely physical, yet via negativa physicalism makes the wrong ruling on them. As I already explained, these are infinitely decomposable worlds with an infinite descent of mental properties, or priority monist worlds with cosmos-spanning mental properties.

These are serious problems for both theory-based and via negativa physicalism, and I do not pretend to have perfectly satisfactory solutions to them. However, I have in mind some some stopgap measures, and hopefully they prove good enough to allow me to move on to a discussion of the specific existentially relevant consequences of the two popular versions of physicalism. In brief, I believe ideal-theory-based physicalism can be made less vague by specifying the essence of scientific inquiry in physics, and that via negativa physicalism can (for now) avoid its problems by bracketing off the classes of possible worlds that it makes the wrong ruling in. Again, though, these are merely stopgap measures.

Starting with theory-based physicalism: the ideal physics horn of Hempel’s dilemma is troubling because we have no idea what ideal physics is like, and if we specify it to simply mean a physics of everything, then physicalism is trivially true—no good! However, both the vagueness and triviality problems of this horn can be avoided if the essence of physics is defined. There are many ways one might do this, but being overly fine-grained is inadvisable since (i) we don’t want to recommit the mistake of materialism and too specifically define the nature of the scientifically-discernable world from an armchair, and (ii) scientific practice is extremely diverse, and a description that is too fine-grained is liable to leave something out.

My suggestions for theory-based physicalism are the following.17

---

17 There might be other ways to go with a specification of ideal physics, but the ones I am aware of have serious flaws. For instance, perhaps Chalmers’ (1996) understanding of ideal physics as quantifying over all and only structural and dispositional properties might work to avoid vagueness and triviality. The problem with this view is that it commits the physicalist to the unlikely position that the universe is nothing but structure and disposition all the way down, with no categorical role-fillers of any kind at the bottom of nature. Notice that this is different than the more popular and plausible view called “dispositional essentialism”, which says that categorical properties have their dispositional roles essentially. The
Law constraint: all events are law-governed in worlds where physicalism is true.

Capacity constraint: there is nothing that is in principle beyond the scientific grasp of beings with the equivalent of human cognitive and perceptual capacities in worlds where physicalism is true.

The law constraint says nothing about the metaphysics of laws of nature: they can be either Humean or non-Humean, and explained modally or otherwise. It is rather a mere commitment to the existence of lawful or law-like regularities in nature which govern all events. These regularities can be fully causally determined by antecedent events—as in classical physics—or underdetermined by antecedent events—as seems to be the case in quantum physics. If the latter, the causally underdetermined events must still be law-governed: events must have well-defined probabilistic outcomes, and cannot be completely random. This gives a sufficient-for-false condition for theory-based physicalism: if there are any non-law-governed events in a world, then physicalism is false in that world.

One might worry that I am reversing the standard way of articulating nomological versus other sorts of possibility. Typically, something is thought to be nomologically possible if it is allowed relative to the lawful constraints that physics specifies. Here, though, I am defining ideal physics relative to nomological possibility, which seems to put the horse before the cart. Is this a serious problem? I think not: by "law-governed" I merely mean that there is some constraint on what can happen in a world which is narrower than metaphysical or logical constraint, and that this constraint is specifiable in a general way. The constraint that e.g. ‘event e cannot occur in place p at time t’ is too specific to count as a law of this sort—constraints which are temporally or spatially indexed to particular times or places are thus ruled out.

The capacity constraint says that an ideal physics relative to human cognitive and perceptual capacities is an account of everything that exists in some world. This is supposed to capture the scientific imperialist spirit of physicalism that there is nothing which is in principle unobtainable by our scientific grasp. It follows from this constraint that if theory-based physicalism is true, then we are not cognitively or perceptually closed from a complete understanding of the natural world. Note that the capacity constraint does not require that there must really be intelligent beings in a world in order for physicalism to be true in that world: the “in principle” in the constraint is there to avoid this conclusion. It is sufficient that all of the structure-only view is not without adherents—e.g. Ladyman (1998)—but it seems unfair to force this view on a physicalist as a matter of definition. The way of cashing out ideal-physics physicalism that I provide is not without problems, but so far as I can see it is the most viable articulation that has yet been given.
things in a world be within the theoretical grasp of beings who are cognitively and perceptual constrained as we are, regardless of whether there really are any such beings in that world to do the grasping.

While there might be some issues, these two constraints help ideal-physics physicalism avoid the vagueness horn of Hempel’s dilemma, since we know what it means for an event to be law-governed, and we seem to roughly know the limit of human cognitive and perceptual capacities (as revealed by cognitive science). Neither is this sort of physicalism trivially true: again, a completely random or uncaused event would make this ideal-physics physicalism false, as would the existence of anything which is in principle not scientifically discernable by beings like us.

The issue with via negativa physicalism is a bit more complex. The problem is that via negativa says physicalism is false if there is fundamental mentality or an infinite descent of mentality, but does not make a distinction between sorts of mentality that are physically acceptable versus sorts that are unacceptable. As I pointed out, there are worlds in which physicalism is plausibly true, yet in which there is an infinite descent of mentality (an infinite descent of functionalist or identity theoretic mentality) or there is fundamental mentality (a priority monist world containing only a brain). I suspect that this problem might eventually be fixed by giving an account of “physicalism-friendly” mentality, which will allow us to distinguish worlds containing physically unacceptable mentality from worlds containing physically acceptable mentality. However, for now I have a less ambitious strategy in mind: bracketing off the problematic possible worlds.

---

18 An obvious issue: there are things that seem physically kosher, yet would make this theory-based physicalism false. For instance, imagine a substance that is in most respects just like the substance which composes humans and stars and planets, but which cannot at all causally interact with us or things like us. Since it would not causally interact with us or our instruments, we would be perceptually closed from knowledge of this substance. Ought the existence of such causally isolated stuff make physicalism false in a world? I am not sure: perhaps it should, if the spirit of physicalism is the spirit of of scientific conquest over all of nature. This is clearly a problem for the view, but I suspect it will remain so for any version of theory-based physicalism.

19 I admit that ‘human cognitive and perceptual capacities’ is a bit vague. In a criticism of McGinn’s (1989) mysterianism about consciousness, Dennett (1991) says: “His thesis about the likely limitations of our brains would be uncontroversially true if it weren’t for our clever trick of expanding the powers of our naked brains by off-loading much of the work to artifacts we have designed and built just for this purpose. The brains we were born with are no doubt quite incapable of grasping long division—let alone calculus or photosynthesis—without the aid of pencil and paper or chalk and blackboard.” In other words, our capacities are obviously enhanced through tool use, and who knows what the upper limit of tool use is? It gets even worse if extended mind (Clark and Chalmers 1998) considerations are brought in, and our capacity-enhancing tools are sometimes parts of our minds. Nonetheless, it seems to me that humans must have some innate cognitive and perceptual limitations, regardless of what tools we employ, and that we can at least roughly discern our own cognitive and perceptual limits.
Rather than refine *via negativa* physicalism, I will leave it as the thesis that there is no fundamental mentality or normativity. Worlds that are infinitely decomposable or in which priority monism is true are—for present purposes—to be bracketed off. In the subsequent assessment of the existential consequences of physicalism, I will not consider worlds in those classes. As I said, this is merely a stopgap measure, but hopefully will do in this context—until someone comes up with an improved version of *via negativa* physicalism, with an adequate explanation of the difference between physically acceptable and physically unacceptable mentality, the present notion is the only one available.

Now, equipped with patched up versions of theory-based and *via negativa* physicalism, I can engage in the project of determining the existentially relevant consequences of physicalism. I will do this by progressively moving through the list of prima facie existentially relevant consequences of physicalism that I earlier identified—that physicalism is inconsistent with the existence of robust free will, God, an immortal soul, ultimate purpose, and normatively-laden laws of nature like karma; and that physicalism entails that we need not fear supernatural powers and can in principle know everything scientifically—in order to determine which of these *really are* consequences of the two notions of physicalism under consideration.

2.1 Free will
There are robust and non-robust senses of ‘free will’—the robust sense corresponds to “libertarian” accounts of free will and the non-robust sense corresponds to “compatibilist” accounts. Here I am only concerned with libertarian accounts, which all take free will to be incompatible with causal determinism, such that any world in which determinism is true is a world which can contain no persons with free will. Libertarians generally think that, at minimum, a person who possesses free will must be able to do otherwise than perform some actions that she actually performs. Compatibilism, which is consistent with causal determinism, and often defines free will in terms of an agent’s capacity to satisfy one or another of her desires (sometimes her “deep” desires—see Frankfurt 1971, Wolf 1987), seems to not be at all problematic under any version of physicalism (assuming that physicalism does not entail radical eliminativism about desires or agents). However, many people intuitively feel that something of existential importance is lost if only a compatibilist notion of free will is true, so here I will examine whether libertarian free will could possibly exist in a world that is entirely physical.
Libertarian theories of free will fall into one of three general categories: noncausal (e.g. Ginet 1989), event-causal (e.g. Mele 1996) and agent-causal (e.g. Chisholm 1976, Clarke 1993). Noncausal accounts do not give a positive causal requirement for free will, but rather say that possession of free will depends on satisfaction of something other than a causal condition (like realization of "actish" phenomenology) along with satisfaction of negative causal conditions (like causal underdetermination). Event-causal accounts combine features of compatibilism—for instance, that freedom requires an action be caused by the person’s deeply held convictions—and add that a free action must be nondeterministically caused. The free agent’s actions must have the appropriate causal relationship to the agent’s beliefs and desires, and the agent must have been able to have done otherwise. Agent-causal libertarianism says that an agent has free will if she is at least sometimes the ultimate causal source of her actions.

I find Derk Pereboom’s (2014) “vanishing agent” argument against non-agent-causal libertarian accounts of free will to be quite persuasive. The gist of it is as follows: if the agent is not the ultimate causal source of her actions, then we cannot hold her responsible for what she does. Rather, non-agent causes would be responsible for the agent’s actions. Libertarian theories of free will are supposed to track responsibility of the agent, so non-agent-causal accounts of free will are not really libertarian accounts of free will at all.

Whether or not this vanishing agent argument ultimately succeeds, I will here focus on agent-causal accounts of free will, since such accounts more obviously express our intuitions about free will in the robust sense. I take it that agent-causal accounts require that the agent be fundamental: if the agent’s actions are fully determined by properties of non-agentive things—like the events that comprise the neural activity of her brain—then the agent is not the ultimate causal source of her actions. Though a decision-making event can be caused by a non-agent thing—e.g. something in the agent’s environment may cause a disjunctive decision-making event ‘decide action A or action B’ to occur—the final event must

---

20 Pereboom is specifically interested in cases involving an agent who has conflicting interests which are in equipoise, such that her e.g. moral desires and self-interested desires are at odds and are perfectly balanced. In these cases, an action cannot be performed on the grounds that it better fulfills whatever the agent’s stronger interests are. On non-agent-causal accounts of free will, the action in these cases will ultimately be caused by something non-agential.
be caused by the agent: causal indeterminacy in a decision event must be ultimately “resolved” (for lack of a better word) by the agent. Is this account of free will consistent with both notions of physicalism?

2.2 Free will and via negativa

Agent-causal accounts of free will seem very clearly inconsistent with via negativa physicalism: agents are mental, and agent-causal free will requires that agents be fundamental, so agent-causal accounts of free will entail that there is fundamental mentality. But perhaps that is too fast—is it possible for an agent-causal account of free will to be satisfied without the agent being fundamental? If so, then there is some possible world in which via negativa physicalism is true and which contains agent-causal free will.

Suppose that the agent is the brain—a high-level, complex system of neurons that does not depend on lower-level mental things. Also suppose that causal indeterminism is true, and further that brain events at least sometimes are causally underdetermined—such that it is not determined in advance whether e.g. brain event A or brain event B will occur at time t. In the actual world, this might happen if brain events directly involve subatomic properties, inheriting causal indeterminacy from those properties (e.g. Penrose 1989). Since the agent is identical to the brain, and the brain is causally underdetermined, does this mean that such an agent might possess agent-causal free will?

No: if the agent is the ultimate causal source of her actions, then causal indeterminacy must be resolved by the agent. Indeterminacy cannot be probabilistically resolved by chance or some other non-agent phenomenon. In the brain case, wherein quantum effects produce underdetermined events, the underdetermined cause resolves into a specific effect at the level of fundamental physics—which is not within the agent’s power to control, unless the agent is also fundamental. Given this, it seems that via negativa physicalism really is inconsistent with the existence of agent-causal free will.

2.3 Free will and theory-based physicalism

Theory-based physicalism seems to be consistent with agent-causal free will. Ideal physics as I have defined it allows for fundamental mentality as well as causal indeterminacy. Combined, these seem sufficient to allow an agent to at least sometimes be the ultimate causal source of her actions—perhaps fundamental agents could exercise ultimate causal responsibility via wave function collapse, or something similar. But doesn’t my version of theory-based physicalism require the existence of laws of nature, and robust free will require that some events not be constrained by natural law?
No: robust, agent-causal libertarian free will does not require that free actions are entirely unconstrained by laws. This view just requires that the agent is sometimes the ultimate causal source of her actions. The decision event can have a probabilistic structure, which may be well defined. Of course, there may be worlds containing agent-causal free will in which decision-making events are entirely uncaused, and not limited by even probabilistic laws of nature. Theory-based physicalism cannot be true in those worlds, but this does not mean that no theory-based physical world contains agent-causal free will. So theory-based physicalism is at least consistent with agent-causal free will—and thus is not existentially terrifying in this regard.

3.1 God
Across history and culture, there have been a staggering number of different ideas about the nature of God—even within Christianity, notions of God are diverse. I cannot consider all such notions of God; for present purposes I will understand ‘God’ to refer to an all-powerful, all-knowing, perfectly good being of pure spirit, who created the universe. Of course, given the diversity of views about the nature of God, this notion of God will not satisfy everyone. For instance, some believe that God is a necessarily existing being (Kripke 1980), or that God exists outside of space and time (Hart 2013). However, a notion of God without these exotic properties is, I think, robust enough to count as God by most people’s lights.

Additionally, there have been thousands of years of philosophical and theological debate about how the properties of God ought to be properly understood. My interest is in determining whether a being who bears all of these properties could exist in an entirely physical world, but not every property of God is relevant to this interest. Here it will suffice to gloss the precise meanings of these terms, unless one appears to be inconsistent with either via negativa or theory-based physicalism. It goes without saying that many people find existential comfort in God, and, as Nietzsche (1887/1974, §125) says, would find themselves “plunging continually … [b]ackward, sideward, in all directions” as if lost in an endless expanse of space, if deprived of the reassuring presence of a divine being.

3.2 God and via negativa physicalism
If God is defined as a being of “pure spirit”, and this means that God is fundamentally mental, then God trivially cannot exist in a world in which via negativa physicalism is true. Since it seems natural to interpret
“pure spirit” as “fundamentally mental”, it easily follows that God could not be physical in the *via negativa* sense.

But this is too easy. Remember that I am interested in the existentially relevant consequences of physicalism. Perhaps what is existentially relevant here is not whether a being of pure spirit exists, but whether a being who instantiates all the other properties commonly attributed to God exists. Suppose theological leaders were to come to the conclusion—and disseminate to their flocks—the idea that God exists, and is the all-powerful, all-knowing, perfectly good creator of the universe, but that God is not a being of pure spirit. Rather, God’s mentality is non-fundamental: perhaps God is a cosmic spacebrain floating around between the galaxies, or a background energy force spread throughout the universe with a mind which depends on the interactions of God’s infinitely many energy fields, or a higher-dimensional being who has a nature that we cannot fully comprehend, but is nonetheless not fundamentally mental. Would anyone find the mere fact that God is physical in this sense to be a reason for despair?

This is a bit tricky. Historically speaking, theists have sometimes gotten rather hung up on seemingly minor details pertaining to their religious beliefs, like the disagreement between Protestants and Catholics on whether to understand the wine in the Eucharist ritual as literally transformed into the blood of Christ. Perhaps there are some theists out there who are personally or ideologically committed to the existence of a God of pure spirit. Perhaps they even feel so strongly about this bit of doctrine to consider its denial heresy, and might even consider life less worthy of living if the doctrine were false.

However, such an attitude strikes me as irrational. It seems that *via negativa* physicalism is consistent with all the other properties of God. That is, the nonexistence of a being of pure spirit in a world does not seem to entail that the world could not contain a being who is all-powerful and all-knowing and so forth. So, while a God of pure spirit could not exist in a world where *via negativa* physicalism is true, it seems that a non-pure-spirit God could exist. And it is these other properties of God that a person might reasonably derive existential significance from: God as creator and maintainer of the world’s moral order, who pays attention to and loves all of us. It seems that a person should derive just as much comfort from
the existence of a being who can do all this, yet who is not a substance of pure spirit. The “pure spirit” bit, in my eyes, does no real existential work.²¹

However, one might object that there is a tighter relationship between the various properties of God than I have been suggesting. Perhaps God could not be e.g. perfectly good or the creator of the universe unless God is also a being of pure spirit. Let’s look at the details of how this response might work.

Does God-as-creator depend on God-as-fundamentally-mental? Here is a reason to think so: the cosmological argument for the existence of God roughly says that all physical things have a cause. I take it this is because it is inconceivable for there to be an endless backward chain of causes without beginning (put aside the plausibility of this premise and take it as a supposition of the argument). So the universe must have been caused by something nonphysical: a being of pure spirit, under the via negativa reading.

The word “physical” clearly does lots of work in this argument. If the main premise were to simply say that all things full-stop must have a cause, then we would need to explain the existence of God as well. God’s nonphysicality offers an out: since God is a fundamentally different sort of thing than the physical universe, we can explain God as always existing or being self-creating, which are (apparently) not options available to explain the origin of the physical universe.

Unfortunately, this argument is quite obviously ad hoc in a bad way: why should it make any different whether God is fundamentally mental or not to the sort of explanation we can give for God’s origin versus the origin of the universe? It doesn’t look to me to make any difference whether something is fundamentally mental relative to its capacity to have always existed or be self-creating. Perhaps the move appeared more compelling to the theists of the past because “physical” and “pure spirit” were

---

²¹ This points to an ambiguity inherent to my overall project: it is not clear whether ‘existential condition’ is merely descriptive, or whether there is an additional normative component. If the former—for instance if the term just refers to what some population finds existentially significant—then I cannot criticize theists who have an existential commitment to the existence of a being of pure spirit. If the latter, then it seems I must provide a more robust account of which existential concerns are reasonable and which are not—a project which I doubt could be satisfactorily accomplished. While I acknowledge this as a problematic dilemma, and I don’t have a solution, I nevertheless don’t think it damns my whole project. I am mostly using ‘existential condition’ in the purely descriptive way, which seems unproblematic. When I venture into normative territory, as in the discussion of a God of “pure spirit”, I am relying on normative intuitions about what is a reasonable existential concern, rather than a more substantial theory of the normative side of ‘existential condition’.
underdefined. However, once the ambiguity is removed, it becomes clear that there is no significant connection between being the creator of the universe and being fundamentally mental.

What about God’s omniscience? One might think that in order for God to know everything, God must have a mind that can store an infinite amount of information, and that no physical vessel can suffice. Only a fundamentally mental storage medium, which is not limited by constraints of the non-mental universe, could suffice.

Again, upon inspection this falls apart. There is nothing in the notion of fundamental mentality which says anything about information storage capacity. A fundamentally mental being might have a limited storage capacity, and a non-fundamentally mental being might have an infinite storage capacity (as in an infinite universe, for instance). So there is no conceptual link between omniscience and fundamental mentality.

Does God need to be a being of pure spirit in order to be perfectly good? It seems not: one might act in the best possible way regardless of whether one possesses fundamental mentality or not. However, here it gets a bit trickier for via negativa physicalism, since this is additionally the view that there is no fundamental normativity.

Could God be morally perfect if there are no fundamental norms? This depends on what “morally perfect” means. If “morally perfect” is defined relative to anything non-fundamental, like a set of social conventions or perhaps rules built into rational thinking, then of course God could be morally perfect in a world in which via negativa physicalism is true. But if a being can only be morally perfect if norms are fundamental, then God could not exist in this world. I’ll note two things: (i) the commitments of via negativa can in principle be broken off from one another, such that there might be a version of via negativa which is consistent with God and another which is inconsistent; and (ii) this is a tricky area of metaethics upon which (so far as I can tell) there is little-to-no agreement. So staking the inconsistency of via negativa and God upon the via negativa inconsistency with fundamental norms may not be wise for a via negativa physicalist of the atheist persuasion.

Finally, how about God’s omnipotence? Need God be a being of pure spirit in order to be all powerful? Here is a reason to think “no”: a world that does not contain fundamental mentality need not be constrained by anything like the laws of physics. God might exist in a world in which God can do anything
that is metaphysically possible, which seems to be consistent with via negativa. However, there are two wrinkles with this.

One wrinkle is that fundamental mentality and fundamental normativity cannot exist in a world in which via negativa physicalism is true, so God seems to be prevented from creating fundamental mental or normative things. However, it is not quite true that this is a limit on God’s power: perhaps God could create Cartesian souls, but decides for whatever reason not to. In such a case, the via negativa constraint on physicalism is not broken—via negativa physicalists typically admit that physicalism is a contingent thesis, and that it is possible for physicalism to be false in worlds which add the wrong things to what exists in our world. The capacity of God to create such “wrong” things seems not enough to make via negativa false in that world as via negativa is normally stated, but it is worth thinking about whether the mere capacity for a world to contain fundamental mentality is sufficient for via negativa to be false in that world—perhaps a modification to the view is in order.

The other wrinkle has to do with what I already said about free will. I said that robust free will is inconsistent with via negativa physicalism: robust free will is agent-causal free will, and agent-causal free will requires fundamental agents. Now: if God is all powerful, then shouldn’t God be capable of acting with free will in the robust sense? And if so, doesn’t God’s omnipotence require that God be a fundamental agent, thereby violating via negativa?

This is a bit thorny. As I pointed out, via negativa is in principle consistent with anything whatsoever happening in a world, except the creation of fundamental mentality or normativity. So (nearly) any of God’s desires can be satisfied without violating via negativa. By my standards, this suffices to make God all-powerful and free, but perhaps I have compatibilist intuitions that are muddying things up. Since I admit that robust free will is inconsistent with via negativa, it follows that—if you think that God’s omnipotence requires free will of this sort—a fully all-powerful God could not exist in a via negativa world. Nonetheless, it looks to me like this is not sufficient to rule that God (as typically thought of) is inconsistent with via negativa physicalism: God can still be the creator of the universe, perfectly good, know everything, and capable of causing any event to occur (except creation of fundamental mentality or normativity).
Thus, even upon further reflection, it looks like *via negativa* physicalism does not have the negative existential consequence that God could not exist. The truth of *via negativa* would merely mean that only a slightly less robust sort of God could exist than the one that theists normally consider—but such a physical God is one who could plausibly satisfy all of the existential concerns which the full-blooded, pure spirit God satisfies.

3.3 God and theory-based physicalism

Could God exist in a world containing only the posits of an ideal physics? I think the answer is plausibly “no”: various of the properties I have attributed to God are strongly in tension with theory-based physicalism as I have described it. Remember that theory-based physicalism has two commitments: (i) the universe is law governed and (ii) ideal physics is constrained relative to human cognitive and perceptual capacities.

If God can do anything, then God can act in a non-law-governed way: God could do things that would otherwise be prohibited by the normal laws of nature, and we would not be able to predict what would happen during those events, even probabilistically. This conflicts with (i). Further, if God knows everything, then God might plausibly have knowledge that surpasses human understanding—God might have e.g. mathematical knowledge that is outside our ability to comprehend. And again, as an all-powerful being, God might be capable of performing actions that are beyond human understanding. These considerations violate (ii).

There may also be an additional issue with God as the ultimate causal source of the universe: perhaps we are cognitively closed from understanding how God could create the world from nothing. On top of this, it seems that a world-creating event cannot be constrained by the laws of a world, since the world’s laws only exist once the world has come into existence. Thus, many of the core properties of God seem to be in tension with theory-based physicalism.

I think these many tensions are irresolvable: God could not exist in a world in which theory-based physicalism is true. Unlike *via negativa* physicalism, the properties of God that are in tension with this version of physicalism are the properties that are existentially significant: God as an all-powerful and all-knowing creator of the universe. So, while both notions of physicalism I am considering are strictly
speaking inconsistent with the existence of God, only theory-based physicalism has the full negative existential significance of atheism attached to it.

4.1 Immortal soul

I will interpret “immortal soul” in a Cartesian way: this is a mental thing that a person is identical to and which interacts with the body to cause behavior. An immortal soul cannot be destroyed once it has come into existence. If one thinks that personal identity depends on certain psychological properties of a person—like her personality or memories—then the immortal soul must also be the vehicle for personal-identity-determining psychological properties, since the person is identical to the soul. Obviously, many religious people are existentially interested in the existence of an immortal soul, such that they would likely feel that life is less worth living if the soul does not exist.

4.2 Immortal soul and via negativa physicalism

This is looks like the easiest existential consequence of physicalism to figure out: via negativa physicalism is straightforwardly inconsistent with the existence of an immortal soul, since souls are fundamental and mental. However, as in the God case, there is a further question: could something like a soul exist, which satisfies the same existential concerns as that of an immortal soul, yet which is not a fundamentally mental thing? As with God, I think the answer is “yes”.

Imagine some persons in a computer simulation: the world as they know it, and they themselves, are computer programs running on digital computers. Suppose that when a person in this virtual world “dies”—as when her “body” is “destroyed” in the virtual environment—her mind is not destroyed. Instead, a new avatar body is created for that mind, which the person uses to continue to engage with the virtual world. The persons in this simulation world are effectively immortal until the computer is turned off—and if the computer is never turned off (which is a possibility consistent with via negativa physicalism), then they live forever.

These digital people do not have immortal souls in the Cartesian sense: their mentality is not fundamental, but rather the product of sophisticated operations performed by lower-level computational hardware. Though they do not truly have immortal souls, are the same existential concerns satisfied as those which would be satisfied by possession of immortal souls? I think so.
From the perspective of these digital people, nothing would be different if a fundamental soul were doing the work to keep them in existence. The realizer for their minds—computer hardware, in this case—continues to exist even after the destruction of their perceptible “bodies”. These persons need not fear death of the body. This, I think, is what is really existentially at stake with regard to an immortal soul. Thus, though via negativa technically is inconsistent with the existence of an immortal soul, the view nonetheless does not have a corresponding negative existential weight. This is because we can satisfy the soul-concern even if we do not properly have souls.\textsuperscript{22}

4.3 Immortal soul and theory-based physicalism

Could an immortal soul exist in a world in which theory-based physicalism is true? I think so: such a world could last forever, and there could be fundamental mentality in that world. Souls could operate in a lawful way, and the way they work could be completely transparent to human scientific scrutiny—thus satisfying the two conditions I posited for theory-based physicalism. So, as I see it, theory-based physicalism is straightforwardly consistent with the existence of immortal souls, and people existentially concerned with the existence of an immortal soul need not fear this form of physicalism.

5.1 Ultimate purpose

The topic of “ultimate purpose” is somewhat more nebulous than the others I’ve so-far discussed: analytic philosophers have not devoted a considerable amount of effort to determining what it means for a person or the universe to have ultimate purpose. The small amount that has been written on this is largely religious, such that ultimate purpose is derived from the dictates of God (e.g. Affolter 2007) or from the essential nature of the soul (e.g. Tolstoy 1884/2010, Craig 1994). Here I will assume that ultimate purpose need not be determined by religious entities of this sort.

By “ultimate purpose” I mean the following: someone or something has ultimate purpose if (i) the person or thing ought to have some property or properties, (ii) that property or those properties confer positive value to the possessor, and (iii) the “ought” does not ground in anything more fundamental. In other words, if there is a way that a person or thing should be in order to be valuable, and this normative

\textsuperscript{22} One might worry that via negativa seems to require that we exist in a computer simulation in order for us to have souls, which might be independently existentially unsettling. However, this is not the case: a similar situation is possible if our minds were stored on indestructible vehicles, which grows or gets attached to a new body after the death of its last host body. And there are surely other ways to realize pseudo-immortal souls. So via negativa does not have the negative existential consequence that if we have immortal souls (or pseudo-souls), then we are denizens of a computer simulation.
fact is not explainable nor reducible to anything non-normative (like a non-normative evolutionary or social fact), then that person or thing has an ultimate purpose. This means that subjectivist accounts of the meaning of life, for instance Frankfurt (1982), do not count as providing an account of the ultimate purpose of life, since these accounts typically deny ultimate human-psychology-independent facts about normativity or value. Only “objectivist” (Metz 2013) accounts of life’s meaning or purpose, such as in Wolf (1997), Nozick (1981), or Railton (1984), seem consistent with the criteria I have provided.

Notice that under this definition, not every value-conferring property is sufficient to confer ultimate purpose. Suppose that having a pleasurable life is value-conferring, but that there is no fundamental norm that dictates a person ought to live a pleasurable life. In this case, having pleasure may make a person’s life better, but not imbue it with purpose. Alternately, if it is true that a person ought to live in accordance with her true nature, or ought to be a moral person, and these properties confer value on the person, and the ought claim is not explained by more fundamental non-normative properties, then these properties are sufficient to give a person ultimate purpose.

5.2 Ultimate purpose and via negativa physicalism

As I have defined it, ultimate purpose requires that fundamental norms exist: if the norm that a person ought to have property P depends on a non-normative property N, then P cannot provide ultimate purpose to a person. Though I have not yet discussed it in any depth, via negativa physicalism has a commitment to the nonexistence of fundamental normative properties. Given that, via negativa seems outright inconsistent with the existence of ultimate purpose for individual lives or the universe as a whole.

There is no existential wiggle room for via negativa physicalism on this, unlike the case with God or the soul. What is existentially relevant to many people, I take it, is that there is ultimate purpose, which is purpose that has some fundamental positive normative aspect. Tasks which function to attain some intermediate end (like manual labor, perhaps), or things we find intrinsically valuable but which do not satisfy the norm that we ought to attain them (like pleasure, perhaps) are not ultimately purposeful to us. Ultimate purpose instead requires a deeper connection to what is fundamentally and intrinsically good and worthwhile, which requires the existence of fundamental normative facts or properties. Imagine finding out that all of our desires were implanted into us by ancient aliens who return to Earth every few millennia in order to feed on the byproducts of our civilization (e.g. plastic), and that this is the ultimate
explanation for all of our interests—there are no deeper normative facts upon which our interests depend. I suspect that in this scenario, interest-based facts would fail to satisfy the existential concerns that are satisfied by possession of genuine ultimate purpose. Given these considerations, via negativa physicalism as I have defined it is inconsistent with anything which could fulfill our ultimate-purpose-based existential interests.

However, one could remove the no-fundamental-normativity constraint from via negativa, which would allow ultimate purpose to exist. While this is a perfectly acceptable way to modify via negativa—and is largely consistent with via negativa’s philosophy of mind roots (though rationality is often considered essentially normative)—I am here only interested in the full-blooded version of via negativa, with all of its attendant existential consequences. I have two reasons: to limit the scope of my project (I cannot entertain every conceivable version of physicalism), and because I think full-blooded via negativa captures more of our intuitions about the “spirit of physicalism”: full-blooded via negativa seems to capture a Copernican spirit of not positing any human-specific properties at the fundamental of reality, whereas thinner (no-fundamental-mentality-only or no-fundamental-normativity-only) notions don’t do as good a job at capturing this spirit.

5.3 Ultimate purpose and theory-based physicalism
So far, I have been assuming that neither notion of physicalism entails the other notion, but perhaps this is a mistake. This assumption is relevant at this juncture because if theory-based physicalism allows for fundamental normative facts or properties, then it is apparently consistent with the possibility of ultimate purpose. But if theory-based physicalism entails that fundamental norms cannot exist, then it is—like via negativa physicalism—inconsistent with the possibility of ultimate purpose. Does theory-based physicalism entail that there is no fundamental normativity?

I take it that the logical positivists believed science is incapable of discovering or even describing fundamental normative facts, due to the nature of science as the project of discerning and cataloguing regularities in sense-data. This, combined with the verificationist thesis that the only meaningful non-analytic and fact-stating sentences are ones that can be empirically verified, led to the development of non-cognitivist views such as emotivism, which says that normative assertions are not fact-stating, but rather that they are emotional expressions of our preferences for how people ought to behave or what
they ought to desire, and those preferences are determined by non-normative psychological properties. Under this view, the utterance ‘murder is wrong’ really just means ‘murder: boo!’.

Were the positivists right? Can science not possibly describe the world as containing fundamental normative facts? Consider the two commitments of theory-based physicalism: that all events are constrained by natural laws, and that everything that exists is in principle describable by an ideal human physics. The first constraint seems not to be violated by the existence of fundamental norms: the existence of fundamental norms does not require that there be events which are not law-governed.

What about the ideal human physics constraint? This is a bit loose, but it is nonetheless intelligible to ask whether human cognitive and perceptual capacities prevent us from ever having a science that could describe the world as containing fundamental norms. To start: if we were incapable of conceiving of fundamental norms, then obviously an ideal human science would not be able to describe the world as realizing those sorts of normative properties. But we seem to be capable of properly conceiving of fundamental normative properties—much of the metaethical literature about moral facts and non-naturalism would otherwise be rendered incoherent. If this were so, then even the paragon defender of common sense, G. E. Moore (1903), would count as not just wrong, but unintelligible!

What about human perceptual constraints? It seems plausible that we in principle cannot perceive normative properties—doing so would violate Hume’s (1738) is-ought distinction, on the assumption that we can only directly perceive non-normative states of affairs. Does this mean normative properties cannot figure into our best science? I think not: we cannot perceive mathematical or other abstract properties either. Yet I doubt very much that many people think this is a good enough reason to believe that mathematical properties cannot be part of our best science. Likewise, though we cannot perceive fundamental norms, this does not mean that they cannot be part of an ideal human science. Hilary Putnam (1982) even argued that scientific practice is inextricably interwoven with normativity—as he put it, there is no fact/value divide. Whether or not this is right, it nonetheless seems that theory-based physicalism is consistent with the possible existence of fundamental norms, and thus is consistent with ultimate purpose for individual humans or the universe.

6.1. Karma

---

23 See companions in the guilt arguments in metaethics, as in Lillehammer (2013).
I said that there is little discussion in analytic philosophy on the notion of ‘ultimate purpose’—well, the same goes tenfold for discussion in analytic philosophy on normatively-laden laws of nature like karma. Nearly all literature pertaining to karma or the like in analytic philosophy is confined to scholarly work on Asian religion and philosophy, though of course this work is not without philosophical merit (see Siderits 2007, Flanagan 2011 and Goodman 2014). And as far as I can tell, no one has weighed in on whether it is possible for a fully physical world to have karmic laws in it, unlike every other topic I’ve discussed so far.

I’m not specifically interested in karma, but rather the class of laws of nature of which karmic law is a member. The laws in this class are causally-efficacious and normatively-laden. These laws are sensitive to human action, and cause states of affairs to obtain in response to an action depending on how that action relates to a norm built into the law. Karmic laws in this class punish or reward actions that respectively violate or satisfy some norm. Though I’m not specifically interested in karma, I will center my discussion on karma as a archetypal law from the class that I am interested in. Though existential concern for karma is probably not as common in America, Europe or Australia (i.e. the current geographic centers of analytic philosophy) as it is in other parts of the Earth, I nonetheless suspect that many people have an existential interest that karmic or similar laws obtain in our world. But could such laws obtain in an entirely physical world?

6.2 Via negativa physicalism and karma

If karma is a fundamental law of nature, and it essentially involves normativity, then karma puts normativity into the fundamental level of reality—thus violating via negativa physicalism’s constraint against fundamental normativity. I will suppose that karmic law is essentially normative. Need karma be a fundamental law?

Many philosophers have become convinced that there are non-fundamental laws of nature, for instance the laws posited by the special sciences which govern non-fundamental phenomena like chemistry and biology. These laws are necessitated by fundamental laws, but are not identical to any fundamental laws. Perhaps karma could be like a law of the special sciences, and thus plausibly non-fundamental.

As in my discussion of ultimate purpose, I will again assume a robust version of via negativa which includes both no-fundamental-mentality and no-fundamental-normativity requirements.
However, there is a difference between karma and special science laws which plausibly prevents karma from counting as non-fundamental. Special science laws (assuming they are irreducible to lower-level laws) pick out real patterns in nature which are not explicit at the level of fundamental physics. For instance, Fodor (1974) provides the example of Gresham’s law in economics that “bad” currency (which is currency which has a nominal value higher than its commodity value) drives out “good” currency (which is currency which has a roughly identical nominal and commodity value) from a market. Gresham’s law provides testable predictions, which have been well demonstrated empirically, yet it is hard to imagine what fundamental laws Gresham’s law might be reduced to. It seems to pick out real patterns at a high level of nature which are not explicitly present at the level of physics.

Real and irreducible as Gresham’s law may be, whatever states of affairs it predicts to come about will be the same states of affairs as what the fundamental laws of physics predicts will come about. This point is demonstrated by Dennett (1991b): hypothetical Martians with only perfectly accurate knowledge of fundamental physics—and no knowledge of higher-level phenomena—could accurately predict what future states of affairs will obtain, even though they do not possess knowledge of real patterns of nature that are described by the special sciences. The special science laws add nothing causally novel to the world: the causal powers (if any) of the properties that figure in special science laws are derived from the causal powers of the properties that figure in lower-level laws.

Karma, unlike special science laws, must involve properties with causal powers that are not derived from the causal powers of properties of lower-level laws, or else karmic properties must depend on fundamental properties that involve normativity. This is because karma is responsive to human action—which is not especially problematic, as Gresham’s law too involves properties that are responsive to human action—and karma causes states of affairs to come about depending on whether the human action violates a norm or not. This norm-sensitivity requires that karma has causal power that does not depend on non-normative fundamental properties—if fundamental laws are not norm-sensitive, then it seems plausible that the higher-level laws that depend on them will not be norm-sensitive either.25

25 One might recall Donald Davidson (1970) famously arguing that the rules governing rationality are both normative and not reducible to lower-level laws, even though they supervene on lower-level laws. If so, then don’t the rules of rationality count as a compelling example of irreducible high-level laws that are essentially normative? I am not at all an expert on Davidson exegesis, which is difficult, but my understanding is that he took rules (which are explicitly not natural laws, since they grant exceptions)
So it seems that *via negativa* is inconsistent with karma, as well as inconsistent with the laws in the class which karma belongs to. However, could our existential concern regarding karma be satisfied by something other than a genuine law of nature? I think the answer is “yes”—just as something other than a genuine immortal soul could satisfy our soul-based existential concerns, something other than genuine karma could satisfy our karma-based existential concerns.

Why might someone find karma to be valuable? I think for the following reason: if karma is real then there is something in the world which ensures that bad actions are punished and good actions are rewarded. Karma creates practical consequences for moral or immoral behavior, and maintains a moral order in the universe (as God does for theists). Such a moral order might be maintained, though, by something other than a full-blooded law of nature.

Again, imagine a population of digital beings who exist in a computer simulation. This virtual world has ersatz laws that are analogous to the laws of physics, as well as an ersatz law analogous to karma. This means that good actions in the virtual world are rewarded, and bad actions punished. From the perspective of these digital beings, there is a law of karma—even though no such law really exists in that world.

My feeling is that the karma-based existential concerns of these people ought to be satisfied, even though karma doesn’t properly exist in their world. If this is right, then even though *via negativa* physicalism is strictly speaking inconsistent with karma, our karma-based existential concerns might still be satisfied in a world where *via negativa* physicalism is true.

6.3 Theory-based physicalism and karma

There seems to be no problematic tension between theory-based physicalism and the possible existence of karma. Karma is a law, and plausibly not principally outside the scope of potential human scientific inquiry. These are the two constraints I have deemed to constitute theory-based physicalism, so karma could exist in a world wherein theory-based physicalism is true.

7. Epicurean freedom and scientific imperialism

So far I’ve only assessed the apparent negative existential consequences of physicalism, which are the consequences of physicalism that plausibly would make a large number of people believe that life is less governing rationality to be causally explanatory, but not independently causally efficacious. This is unlike karma, which has causal power that is independent of the causal power of fundamental laws.
worth living. However, physicalism also seems to have a few positive existential conditions: (i) we need not fear supernatural forces of nature, which is a condition I’ll call "Epicurean freedom" and (ii) we can in principle know all of nature, which is a condition I’ll call "scientific imperialism". If the negative existential consequences of physicalism constitute the horror of physicalism, then these two positive consequences represent the hope of physicalism.

Epicurean freedom inverts what I take to be the typical attitude toward physicalism’s negative existential consequences: some people may think that life is more worth living if God, immortal soul, robust freedom, ultimate purpose or karma do not exist. For instance, atheists may be drawn to the attitude Rorty describes as not needing to justify ourselves before a divine authority. With the absence of ultimate purpose, some individuals may feel more agency to create meaning and purpose for themselves. Someone might find a normatively-laden law of nature like karma to oppressively impose a moral order on the world. Without robust free will, one might feel relieved that humans are not ultimately causally responsible for their actions, and thus perhaps not blame-worthy for their morally bad actions. And like Epicurus, some may be comforted by the idea that life is finite rather than infinite, and that no torments potentially await in an afterlife, or in an endless succession of future lives.

Since Epicurean freedom is just an inversion of the negative existential attitude toward the consequences of physicalism that I have discussed so far, this attitude attaches itself to the two versions of physicalism to the same degree that they entail those various conditions. Given that the same set of conditions can be taken as either existentially positive or negative, one might wonder whether they ought to be considered as making life more or less worth living. What is the correct existential attitude one ought to have toward these conditions?

In my view, this is where reasons end. Some people are so psychologically constituted as to find these conditions to be existentially negative, and some people are constituted to find them existentially positive. Whether one is emboldened or frightened by these consequences of physicalism (in its two guises) largely depends, it seems to me, on how much metaphysical reassurance one requires. Those who can say “yes” to life even in the absence of life-affirming myths may feel emboldened by a metaphysically stark and reassurance-free physical world, whereas those more in need of the comfort of myth will grow terrified.
Scientific imperialism is more straightforward: *via negativa* physicalism obviously does not entail that we can in principle come to possess a complete scientific understanding of the world, and theory-based physicalism does. This is because scientific imperialism is part of the definition of theory-based physicalism, and not part of *via negativa*. Even if there is nothing fundamentally mental or normative in the world, there might still be things about the world we cannot know. If theory-based physicalism as I have defined it is true, then there is nothing in principle that we cannot know through scientific inquiry—though this of course does not mean that we will actually achieve such knowledge, for instance if our species first destroys itself or we stop pursuing scientific knowledge.

8. Conclusion

I've looked at various prima facie existentially relevant consequences of physicalism with the intent of determining which, if any, *really are* consequences of physicalism. Since there are two popular ways to understand physicalism, the project has been bifurcated into determining which are the consequences of theory-based physicalism, and which are the consequences of *via negativa* physicalism. In summary: *via negativa* physicalism entails that there is no free will or ultimate purpose in the world, and is strictly speaking inconsistent with the existence of God, immortal soul or karma. However, it is consistent with facsimiles of God, immortal soul and karma which may be existentially equivalent to the genuine articles.

Theory-based physicalism is inconsistent with the existence of God, but consistent with the existence of immortal soul, free will, karma, and ultimate purpose. Our scientific imperialist urges are satisfied by theory-based physicalism, but not by *via negativa* physicalism. Epicurean freedom, which is a positive existential stance toward the nonexistence of God, soul, free will, karma or ultimate purpose, attaches itself to the two versions of physicalism to the same degree that these entail the nonexistence of these various things—thus *via negativa* seems to gives us more Epicurean freedom than theory-based physicalism. Where does this leave us?

We should endeavor to believe the view which has the best claim to truth, and go where the arguments lead us. Even if someone is inclined toward one or another version of physicalism, with its attendant particular existential consequences, this is not a guide to determining that it is true. For that, we cannot simply look at the existential consequences that I have been discussing, but must look elsewhere. Perhaps one’s preferred definition of physicalism is inconsistent with the existence of God. Well, if there is
some excellent argument that God exists, then this version of physicalism must be given up (of course, I do not think there is any such argument). In the next section I'll consider reasons to prefer via negativa physicalism over other views.
IV. A Minimally Mysterian Solution to the Mind-Body Problem

The causal closure argument, as discussed by Papineau (2001), provide a strong reason to rule out interactive Cartesian-esque dualism: starting in the 20th century, we have acquired compelling empirical evidence that all human behavior can be causally explained without any appeal to fundamentally mental phenomena. To put the point in a slightly different way, it seems that neuroscience describes causally sufficient conditions for the occurrence of our verbal, bodily, and other mind-caused behavior, which appears to rule out a soul (or other non-brain thing) from doing causal work in the production of that behavior. If one were to nonetheless maintain that the mind is completely distinct from the brain, it seems that one would be forced to embrace epiphenomenalism—which says that mental properties are not causally efficacious—or overdetermination—which says that physical effects have both sufficient physical and nonphysical causes. Neither option seems plausible. This constrains the range of solutions to the mind-body problem to some familiar sort of physicalism (I will focus on “Type-B” physicalism, to employ Chalmers’ 2002 terminology) or some variety of Russellian monism (specifically: panpsychism, panprotopsychism, or Russellian physicalism), since these are the only candidate views which are consistent with causal closure and which avoid overdetermination or epiphenomenalism.

The goal in this section of the dissertation is to see which of these views has the strongest claim to the truth. Since these are all views on the nature of consciousness, my focus is on how well these competing accounts explain consciousness relative to each other. But what is consciousness?

Phenomenal consciousness consists of the “what-it-is-likeness” (Nagel 1974) of subjective experience, as typified by one’s feeling of pain in the lower back, the sight of a neon yellow sign, or the taste of well-aged cheese. Consciousness, to put it mildly, is weird: unlike apparently all other natural phenomena, it seems to not be explicable in scientific terms (Levine 1983) nor inferable from scientific facts (Jackson 1982). Yet—putting aside so-called “illusionism” (Frankish 2016) which completely denies the existence of phenomenal consciousness—we know that consciousness is real. It is what we are most and best familiar with, present during all of our waking hours and many of our sleeping ones as well. How

26 I will use ‘phenomenal’, ‘subjective’, and ‘conscious’ interchangeably throughout the text. This is unproblematic because I am not here concerned with non-phenomenal disambiguations of ‘conscious’, for instance access consciousness or report consciousness (see Block 2002).
can something so familiar be so inexplicable, an oddity that is seemingly unique in its resistance to any sort of reductive scientific analysis?

Physicalists generally say that consciousness is not fundamental, and instead typically identify phenomenal properties with neurobiological or functional properties. These philosophers believe that consciousness is not metaphysically special in any way that would substantially distinguish it from other phenomena described by the natural sciences. Nonphysicalists—a group which includes dualists, idealists and panpsychists—instead assert that consciousness is fundamental, and cannot be explained in terms of any lower-level nonmenal phenomena.\(^{27}\)

In my eyes, both of these well-explored stances fail to properly account for consciousness. The traditional physicalist position just described seems committed to the possibility of an explanation of consciousness in terms of scientifically scrutable phenomena, which runs against the common intuition that something important—*the* important something, in fact—is left out of such an explanation. The strangeness of consciousness is seemingly ignored under such an account—no good!

Turning to the other side, nonphysicalism gives up far too easily in the face of the apparent inexplicability of consciousness. While it is perhaps true that phenomenal properties cannot be inferred from knowledge of scientifically scrutable properties (which are properties explicitly referred to by science), and that scientifically scrutable properties fail to explain consciousness, it seems a misstep to arrive at the conclusion that mentality is therefore a fundamental phenomenon. Such a view violates the Copernican attitude of science-respecting and secular philosophers that mentality—which seems to be distinctive of only a select few highly complex organic beings in our tiny corner of the universe—is not a basic feature of reality. Yet physicalism and nonphysicalism are the only available options, since nonphysicalism is just the negation of physicalism. What to do?

Fortunately, a relatively unexplored view called “Russellian physicalism” (Stoljar 2001; Montero 2010, 2015; Pereboom 2011; McClelland 2013; Brown 2017) offers an account of the nature of consciousness which respects and addresses its strangeness without committing the nonphysicalist error of making mentality fundamental. Russellian physicalism asserts that there is more to the physical world

\(^{27}\) I’m avoiding “neutral monism”, since I agree with Amy Kind (2015) that the view is unstable, and either collapses to physicalism (if neutral properties are not mental or specially related to mentality) of some sort of non-physicalism (if neutral properties are mental or specially related to mentality)
than just what is directly referred to by physics. Physics only tells us about the causal, structural, and relational nature of the world—things in physics are defined by what they do, and how they spatiotemporally and causally relate to other things. One might think that there is something left out of this description: the intrinsic nature of things, which is distinct from any dispositional, structural or relational characterization. Russellian physicalism urges that the underlying intrinsic nature of the physical world, which is inscrutable to direct scientific inquiry, consists of “categorical” properties. These categorical properties are not defined by causal dispositions, structure or relations, yet the dispositional, structural and relational properties of fundamental things depend on them. Even though categorical properties are not part of physics, they are nonetheless physical in the via negativa sense, since they are neither mental nor especially tied to mentality.

But what does this metaphysical view about the inner nature of the fundamental physical have to do with consciousness? The following: phenomenal consciousness depends upon both non-categorical and categorical features of the world. Since the view is physicalist, mentality is not fundamental, but an appropriate arrangement of fundamental categorical properties does yield a non-fundamental categorical property which is phenomenally mental.28 Putting it another way, the brain has both high-level categorical and non-categorical properties, and phenomenal consciousness is identical to one or more of its categorical properties.

In addition to Russellian physicalism, several nonphysicalist Russellian views have also been developed (see Chalmers 2015). Here, though, I'm only interested in Russellian physicalism, which denies that fundamental categorical properties are mental or that they bear any special relationship to mentality. Other Russellian views add little new to the debate: Russellian panpsychism (which says that fundamental categorical properties are mental) or panprotopsychism (which says that fundamental categorical properties are protomental) all continue to violate our Copernican intuitions by putting mind at

---

28 Positing high-level categorical properties which depend on both lower-level categorical and non-categorical properties may require expanding the ontological taxonomy beyond what has typically been discussed in this context, e.g. by Pereboom (2013) or Chalmers (2003). More on this in footnote thirty five.
the bottom of nature. No good! Russellian physicalism, however, offers something that these other Russellian positions don’t.

Recall the two problems I identified that theories of consciousness have so-far faced:

(i) Inscrutability: Consciousness cannot be explained by, inferred from, or reduced to scientifically scrutable properties

(ii) Copernicanism: The fundamental level of nature is not tainted by mentality nor by a special relationship to mentality

Most physicalist theories of consciousness thus far have violated Inscrutability, and nonphysicalist theories have violated Copernicanism. Russellian physicalism is a unique view that can satisfy both requirements. Or so it seems!

A savvy philosopher might stop me here, and inform me that this sort of approach is not novel in philosophy of mind: there are at least two sorts of physicalist views that recognize and try to preserve the strangeness of consciousness. These are referred to as “mysterianism” and “the phenomenal concept strategy”. Both admit that consciousness is weird from our perspective, but deny that this epistemic strangeness entails any metaphysical consequences. Mysterianism has been most famously argued for by Colin McGinn (1989), whereas the phenomenal concept strategy has had more widespread support from physicalist philosophers of mind (e.g. Hill 1997, Carruthers 2001, Papineau 2002, Carruthers and

---

29 Russellian nonphysicalism adds the following to the debate: Russellianism gives nonphysicalists a different option to account for mental causation beyond the standard options available to traditional dualism, which are interactionism, epiphenomenalism or overdetermination. Instead, Russellian nonphysicalists can admit causal closure, yet have nonphysical mentality be causally efficacious in virtue of being the role-filler of fundamental physics. In my eyes, though, this is not nearly as significant as the aid that Russellianism gives to physicalism, since at the end of the day any sort of nonphysicalism will continue to violate Copernicanism.

30 One might think that panpsychism and panprotopsychism do not really violate Copernicanism: these views deny that mentality is a distinctively human feature of the universe, and instead assert that mentality is very common. I think this response misses the point: imagine a world in which pantheism is true, and that world’s cosmic God created humans in her image. In such a world, human-defining properties are not specific to humans, since the whole universe instantiates these properties as well. Nevertheless, this scenario clearly violates Copernican intuitions. Any view that makes human-defining properties metaphysically special, I take it, is one that violates Copernicanism—and this includes panpsychism and panprotopsychism.

31 This leaves out Nagelian monism, as discussed by Daniel Stoljar (2015). Nagelian monism says that there are non-categorical facts that we currently do not have access to, and a science which utilizes those facts could explain consciousness. My feeling is that Nagelian monism fails to satisfactorily address anti-physicalist challenges (see Alter 2016), and it has less widespread support than the two physicalist views that I focus on, so will not consider it further in this dissertation. Of course, this does not mean that the view does not warrant further scrutiny.
Veillet 2007, Balog 2009). What is the difference between Russellian physicalism and these other views, especially considering that all accept that consciousness is epistemically special but not metaphysically fundamental?

My project consists in spelling out the difference between Russellian physicalism and these two other physicalist views, and arguing that of all of them, Russellian physicalism offers the most plausible and satisfying account of consciousness. This is significant because the phenomenal concept strategy and (to a lesser extent) mysterianism have been seen as the most plausible options for physicalist philosophers who are impressed by the strangeness of consciousness.\(^{32}\) Hence, the goal here is not just deeper elucidation of Russellian physicalism, as has already been done several times (e.g. Stoljar 2001; Montero 2010, 2015; Brown 2017), but rather articulation of a positive argument for the conclusion that Russellian physicalism offers the best explanation for the mystery of consciousness. Section one will describe the core commitments of Russellian physicalism and explain how it answers typical challenges to physicalism. Sections two and three will distinguish Russellian physicalism from, respectively, McGinn’s mysterianism and from the phenomenal concept strategy, and explain why Russellian physicalism is preferable to either of these views.\(^{33}\)

1. What is Russellian Physicalism?

Russellian physicalism is a version of a more general view called ‘Russellian monism’, which states that the fundamental level of the world has two features: a non-categorical, scientifically scrutable feature and a categorical, scientifically inscrutable feature. The non-categorical part of the fundamental level consists of dispositional, structural, relational and causal properties as described by physics—these properties are defined by behavior and various relations (especially spatiotemporal relations) between fundamental entities.

---

\(^{32}\) Of course, there are many additional options available for those not impressed by the strangeness of consciousness. For instance, there is Block and Stalnaker’s (1999) view that brute explanatory gaps permeate nature (so the explanatory gap for consciousness is not special), and there are a variety of reductive neurobiological (e.g. Koch and Crick 1990), representationalist (e.g. Tye 1995) and functionalist (e.g. Dennett 1978, Baars 1988) accounts of consciousness. I suspect that the recently popular Integrated Information Theory (Tononi 2007), though sometimes adopted by nonphysicalists (e.g. Mørch 2018), will eventually be properly sorted into the reductive category as well (see Mindt 2017 for an argument in this vein).

\(^{33}\) Note that these are not different versions of physicalism, but rather different approaches a physicalist might take on the nature of consciousness and the epistemic and metaphysical challenges it seems to present.
Such scientifically scrutable, dispositionally/structurally/relationally-defined properties have typically been thought to exhaust the set of fundamental physical properties—after all, ‘physicalism’ is often thought of as the view that there is nothing over and above what is posited by physics. However, Russellian monism adds that scientifically scrutable properties depend on even more fundamental categorical properties. These categorical properties are scientifically inscrutable, and thus not directly referred to by physics, yet they are the role-fillers of the physics-scrutable properties. Under this view, something like an electron has scrutable features—defined by what the electron does—and more fundamental inscrutable features—defined by what the electron *is*, independent of its causal dispositions or structural relations.\textsuperscript{34}

According to Russellian physicalism, phenomenal consciousness depends upon fundamental categorical properties which bear appropriate relations to each other. When fundamental categorical properties are properly related, a non-fundamental categorical property is necessitated, which accompanies some non-fundamental scientifically scrutable property—for instance, a neural or functional property. Phenomenal properties are identical to certain of these non-fundamental categorical properties.\textsuperscript{35}

I have previously argued (2017) that Russellian physicalism is properly a version of physicalism in the sense that it does not posit fundamental mental or protomental properties, yet that the view

\textsuperscript{34} This way of describing Russellian physicalism makes categorical and non-categorical properties distinct, and thus not modally coextensive. That is, there are (at minimum) worlds containing our world’s non-categorical properties without our world’s categorical properties. However, one could wed Russellian physicalism to dispositional essentialism (Bird 2005, 2007), which is the view that categorical properties have their dispositional roles essentially, such that categorical and non-categorical properties cannot be modally separated. For reasons that will become clear when I discuss Russellian physicalism’s answer to the conceivability argument, this dispositional essentialist route is not advisable for a Russellian physicalist, since it considerably weakens her response to Chalmers’ anti-physicalist criticism.

\textsuperscript{35} This view suggests an additional ontological category beyond those identified by Pereboom (2013) in his discussion of relational vs. nonrelational properties. He identifies extrinsic properties, which are relational; absolutely intrinsic properties, which are properties that are not relational and which do not entirely depend on the relational properties of the bearer’s parts; and comparatively intrinsic properties, which are non-relational properties which depend on the relational properties of the bearer’s parts. Fundamental categorical properties are absolutely intrinsic, fundamental non-categorical properties are extrinsic, and non-fundamental non-categorical properties are either relational or comparatively intrinsic. Strictly speaking, this taxonomy does have a place for hybrid properties that depend both on absolutely intrinsic and extrinsic properties of the bearer’s parts: these are a non-fundamental species of absolutely intrinsic properties. However, it is convenient to have a term dedicated to this fourth ontological category: I am referring to these terms as “non-fundamental categorical properties”.
nonetheless offers satisfactory answers to various anti-physicalist demands and challenges in a way that non-physicalists should find reasonable. Here I'll briefly clarify and expand upon this argument.

Crucially, if Russellian physicalism is to truly count as a version of physicalism, the following conditions must be satisfied:

(i) No Fundamental Mentality: fundamental properties are not mental
(ii) No Fundamental Protomentality: fundamental properties are not protomental

The No Fundamental Mentality condition needs no explanation: physicalism as I have defined it requires that there be no fundamental mental properties, which is equivalent to this condition. However, the need for a No Fundamental Protomentality condition is less obvious. Why think physicalism is inconsistent with fundamental protomentality? After all, don’t all non-eliminativist, physicalist accounts of the mind deem mental properties to ultimately depend on fundamental properties—thereby making those fundamental properties protomental in some sense?

To answer, I must distinguish between physically problematic and unproblematic senses of ‘protomental’. Some property is protomental in the unproblematic sense if a mental property depends on that property, but there is not a special connection between the lower-level protomental property and the higher-level mental property. For instance, under traditional type-identity theory (Smart 1959), mental properties are identical to neural properties, which makes chemical properties protomental in the unproblematic sense—neural properties depend on the lower-level chemical properties that they are constituted by. These chemical properties are not specially related to mentality: the same chemical properties are capable of constituting non-mental properties, and are not restricted to constituting neural/mental properties. With the exception of eliminativism, which denies any mentality whatsoever exists, all physicalist theories of mentality posit protomentality in this unproblematic sense—they must, for else mentality would be fundamental, and the view would not be physicalist.

However, there is a sense of ‘protomental’ that is physically problematic, and plausibly the existence of fundamental protomental properties in this problematic sense is inconsistent with the truth of physicalism. Problematically protomental properties are properties which have a special relationship to mentality that goes beyond the typical sort of dependence relationship that physically unproblematic protomental properties bear to mental properties. Kevin Morris (2016) proposes a modal test to determine
whether fundamental categorical properties are problematically protomental or not: if it is possible to have a world that duplicates the dispositional/relational/structural properties of our world, but has different or no categorical properties, and if the only additional difference between our world and that world is a difference in mentality, then fundamental categorical properties are protomental in a problematic sense. Morris contends that the fundamental categorical properties posited by Russellian physicalism fail his test, rendering Russellian “physicalism” a version of physicalism in name only.36

The first step of Morris’ test asks you to conceive of a world that is just like our world as it is described by physics, but which does not contain the same categorical properties as our world does—it contains “swapped” (i.e. different) or “bare” (i.e. absent) categorical properties. Aside from dispositional essentialists, who say that categorical and non-categorical properties cannot be separated in this way, most Russellian physicalists will accept this sort of world as a possibility. The reason they typically accept the possibility of this sort of world is a consequence of the Russellian physicalist’s answer to Chalmers’ (1996) anti-physicalist conceivability argument.

The conceivability argument says that it is conceivable that there could be a minimal physical duplicate of our world which does not duplicate our world’s phenomenal properties. Worlds in this class which lack penomenal properties altogether are “zombie worlds”, which contain humans who act conscious and have the same neural and functional properties as us, but completely lack subjective experiences. If these worlds are conceivable, then (according to Chalmers) they are possible. If they are possible, then physicalism is false, since physicalism requires that phenomenal properties be necessitated by physical properties. So, if the premises are accepted, physicalism is false.

Russellian physicalists generally accept all parts of this argument except for the first premise, that zombie worlds are conceivable. This rejection is not on the grounds that consciousness really is a mundane, scientifically scrutable property of the world that can be inferred from lower-level physical properties just like anything else, as reductive physicalists like Dennett assert; rather, Russellian physicalists deny that consciousness is scientifically scrutable at all.

36 A version of Russellian physicalism could be constructed which would trivially pass Morris’ test: this alternative formulation denies that categorical properties are constitutive of consciousness, and instead identifies consciousness with non-fundamental non-categorical properties. Fundamental categoricals would be posited for some non-mind-based other reason, perhaps to make the world concrete rather than abstract. However, such a version of Russellian physicalism adds nothing discourse on the mind-body problem, so I will not consider it further.
The reason zombie worlds seem conceivable is because, whenever we have so-far thought of zombie worlds, we have been thinking of worlds which duplicate only the scientifically strutable, non-categorical aspects of our world. The categorical aspects of our world are left out of such conceptions. A minimal non-categorical-plus-categorical duplicate of our world could not lack the same phenomenal properties as are realized in our world—and so a more robust conception of the physical world leaves no room for the possibility of zombies. This does not mean that we know what it means to truly conceive of a full physical duplicate of our world, or that we can infer phenomenal properties from physical properties: we do not have robust enough conceptions of fundamental categorical properties, and only have acquaintance with them in virtue of having access to our own phenomenal properties, which depend on fundamental categorical properties. Lacking robust concepts for these categorical properties, we cannot infer phenomenal properties from physical properties, nor explain phenomenal consciousness in terms of physical properties, since we ultimately do not fully understand the physical.\(^{37}\)

For this reason, the Russellian physicalist must accept the possibility of worlds with bare or swapped categoricals that are mentally distinct from our world—allowing these possibilities provides the Russellian physicalist with resources to plausibly deny that we can conceive of full physical duplicates of our worlds which are zombie worlds, as well as explain what we are really doing when we think we are conceiving of such worlds. Now: recall that acceptance of the possibility of worlds with bare or swapped categoricals is a necessary (but not sufficient) requirement in Morris' test for problematic protom mentality: categorical properties are only protomental if a change to the categorical properties without a change to non-categoricals is possible, and the change necessitates a change to mental properties. Since Russellian physicalists accept this, does this mean that they must also accept the existence of problematically protomental properties?

\(^{37}\) This appeal to “robust” concepts, in contrast with “thin” concepts, warrants an explanation. I take a robust concept to be one with enough descriptive content to fix a de re reference (restricting the scope of possible referents to the actual world), whereas thin concepts have enough content to fix a reference to a class, but not to any particular member of that class. Our phenomenal concepts are robust, since we can descriptively refer to individual experiences. However, our concepts of fundamental categoricals are thin: we cannot refer to any particular fundamental categorical properties, though we can refer to the class of categoricals. Of course, there are a host of issues concerning descriptive (as opposed to Kripkean/Millean) de re reference which I would like to avoid—hopefully what I have said here will suffice for the purposes of this dissertation without getting me into too much hot water.
The answer is “no”, but before I explain this “no”, it would be helpful to spell out Morris’ argument more explicitly:

1. Russellian physicalism’s unique response to the conceivability argument requires the possibility of worlds which are categorically-thus-phenomenally different than the actual world, yet non-categorically identical (“scrute-worlds”, for short)
2. If there could be scrute-worlds worlds which differ from our world only phenomenally, then categorical properties have a physically-problematic relationship to phenomenology
3. There could be scrute-worlds which are minimal non-categorical duplicates of our world and which are different only in phenomenal properties
   \[ \therefore 4. \text{Categorical properties have a physically-problematic relationship to phenomenology} \]

As I said, most Russellian physicalists will accept the first premise, because otherwise they lose their unique answer to the conceivability argument. For instance, Russellian physicalists who are also dispositional essentialists, and think that properties possess their nomological profiles essentially (such that categorical and non-categorical properties cannot be disentangled), cannot explain what we are conceiving of when we conceive of zombie worlds. They must instead say that we are making a more serious error, since under this view there are no zombie worlds that are even scrutably similar to our world—and thus Russellian physicalists who are dispositional essentialists are left without an answer to this anti-physicalist argument.

Maybe a Russellian physicalist could deny premise two on the grounds that fundamental protomental properties are never physically problematic—this view would say that only fully mental fundamental properties pose a problem for a physicalist. However, I agree with Morris that properties which exist only to generate mentality seem strongly in tension with a physicalist view that denies the existence of fundamental mentality. To borrow a metaphor from Robert Howell (2009), problematically protomental properties are like a mentality-bestowing chair: placing an otherwise unminded thing like a rock in the chair suddenly gives it subjective experience, without any change in the thing’s structure or composition. Surely the existence of such a chair would threaten the truth of physicalism—and so too for problematically protomental properties more generally.
As I see it, the appropriate place to push against Morris’ argument is the third premise, that there could be scrute-worlds (again, worlds that have different categorical properties than our world, yet are non-categorically identical) which are only different phenomenologically. But how can a Russelian physicalist deny this? Mustn’t they admit that scrute-worlds could exist which are different only in phenomenal properties, since they think phenomenal properties depend on fundamental categorical properties?

No: an escape route presents itself if the Russelian physicalist says that phenomenal properties are not the only sort of non-fundamental categorical properties. Rather, the Russelian physicalist should instead say that our world realizes a great variety of non-fundamental, non-phenomenal categorical properties. On this precisification of Russelian physicalism, a high-level object like a desk has non-categorical properties like hardness and mass, and additionally has categorical properties that are inscrutable to us. What do these properties do? Well, since they are non-phenomenal, they cannot give desks subjective experience (quite a moral relief to me, as I am always banging my knees against my desk or slamming my fist down when I get upsetting news). Instead, perhaps these non-fundamental categorical properties make desks concrete rather than merely abstract. Regardless of whatever the more fine-grained description, non-phenomenal categorical properties must give objects a non-mental inner being of some sort that cannot be revealed through observation or interaction.

How does this move help refute Morris’ argument? In the following way: the Russelian physicalist can now say that scrute-worlds would differ in both phenomenal and non-phenomenal respects. Just as a change to fundamental non-categorical properties would change all the higher-level properties that depend on them, so too would a change to fundamental categorical properties affect all non-fundamental categorical properties. In swapped categorical worlds, both phenomenal and non-phenomenal properties would be different than those in the actual world (since the fundamental categoricals are different). In bare categorical worlds, all non-fundamental categorical properties would be absent (since the fundamental categoricals are absent). Thus, the third premise of the argument above can be plausibly denied: scrute-worlds cannot differ only in phenomenal properties. This allows the Russelian physicalist’s fundamental categoricals pass Morris’ test, making them not problematically related to mentality.
So far I have explained Russellian physicalism relative to the conceivability argument; what about the other standard arguments against physicalism, such as Jackson’s (1982) knowledge argument or Levine’s (1983) explanatory gap argument? Russellian physicalism gives straightforward answers to these arguments: when Mary the brilliant neuroscientist leaves her black-and-white room for the first time and sees red, she genuinely learns about a physical property that she did not have knowledge of when she was in the room. She becomes newly acquainted with a non-fundamental categorical property, on the grounds that her scientific knowledge was only of non-categorical properties. She could not have inferred what phenomenal redness is like from knowledge of neuroscientific properties, since phenomenal redness is not a property of that sort. Likewise for the explanatory gap: it makes sense that scientifically scrutable, non-categorical facts fail to explain consciousness, since consciousness is of a different sort. We cannot explain a categorical property only in terms of non-categorical properties.

There are some other issues facing Russellian physicalism, for instance a version of the combination problem (Chalmers 2017) which says that subjects cannot come about from any arrangement of non-subjects, and a criticism that categorical properties are epiphenomenal (Howell 2015), since non-categorical properties seem sufficient on their own to cause physical effects. Rather than develop my own solutions to these issues, I defer to the adequate answers that have been given elsewhere: McClelland (2013) argues that self-representing systems can form subjects, and that the rest of the issue constituting the combination problem comes from a requirement that subjects have intrinsic properties (an issue which Russellian physicalism is obviously equipped to handle). Alter and Coleman (in preparation) argue that categorical properties, as the role-realizers of fundamental non-categorical properties, can be causally efficacious, even though different categorical properties may have scrutable-identical effects in worlds of swapped categoricals.

This wraps up my description of Russellian physicalism. Now to compare this view against what I take to be the best physicalist contenders for a viable theory of consciousness. Is Russellian physicalism really distinct from mysterianism or the phenomenal concept strategy, and if so, are there good reasons to prefer Russellian physicalism to these other views? On both counts, I think the answer is “yes”.

2. McGinn’s Mysterianism
Mysterianism is the view that human beings cannot solve the mind-body problem on the grounds that we are cognitively closed from acquiring a solution. McGinn (1989, p. 350) explains cognitive closure in the following way: “A type of mind M is cognitively closed with respect to a property P (or theory T) if and only if the concept-forming procedures at M's disposal cannot extend to a grasp of P (or an understanding of T).”

If McGinn is right, then we cannot solve the mind-body problem because we cannot grasp the appropriate concepts that would allow us to solve it. Opponents of mysterianism (e.g. Dennett 1991) typically reject the view because it is what it says it is: mysterianism fully accepts that phenomenal consciousness is completely mysterious to us, and that there is nothing we can positively say about it. A mysterian might venture a guess about why consciousness is so mysterious to us, but this can never be anything more than an unfounded speculation—or else we would understand something substantial about consciousness and its place in nature. Such guesses are like the hypotheses that monkeys might make about why they cannot grasp fundamental physics—woefully inadequate and hopelessly unjustified. McGinn offers such a guess when he suggests that our perceptually-derived concepts are necessarily spatial, but that a solution to the mind-body problem perhaps requires non-spatial representations. This hypothesis about the source of the mystery of consciousness is purely speculative, for otherwise we would grasp something about the nature of consciousness (that it requires non-spatial representations, which seems to suggest that it is a nonspatial phenomenon).

Is Russellian physicalism a species of mysterianism? At a glance, it seems that it might be: Russellian physicalism asserts that there are properties which are outside of our scientific grasp, and that the ungraspability of these properties is the source of consciousness’ strangeness for us. However, upon closer inspection, it turns out that Russellian physicalism is not a version of mysterianism as McGinn is conceiving of it. At best, Russellian physicalism might be considered a cousin of mysterianism, but certainly not in the same nuclear family. This is because Russellian physicalism gives a positive account of the nature of consciousness and its place in nature, and diagnoses the strangeness of consciousness as a consequence of facts about the world, not facts about the limited minds of human beings.

That Russellian physicalism is not identical to McGinn’s mysterianism can be clearly seen when you look at what McGinn explicitly says about his view. For instance, McGinn distinguishes his view from
any sort of constructive account of consciousness. “One form [of solution to the mind-body problem], which we may call constructive, attempts to specify some natural property of the brain (or body) which explains how consciousness can be elicited from it. … The approach I favour is naturalistic but not constructive: I do not believe we can ever specify what it is about the brain that is responsible for consciousness” (McGinn 1989, p. 350). Russellian physicalism identifies a natural property of the brain—a non-fundamental categorical property—as what is responsible for consciousness. This makes Russellian physicalism a broadly constructive account of the sort that McGinn rejects.

A more explicit reason to think that Russellian physicalism is not a version of mysterianism comes in footnote two about panpsychism and panprotopsychism. “Attributing specks of proto-consciousness to the constituents of matter is … extravagant. I shall here be assuming that panpsychism, like all other extant constructive solutions, is inadequate as an answer to the mind-body problem” (McGinn 1989, p. 350). Russellian physicalists should agree with McGinn that attributing mentality or (problematic) protomentality to the fundamental level of reality is not a good move. However, they should additionally deny that explanations of this sort aren’t useful: Russellian physicalism is in the same boat as panpsychism and panprotopsychism when it comes to the sort of explanation of consciousness it is giving. All of these views say that positing fundamental categorical properties with intrinsic natures that are not revealed through physics can help us solve the mind-body problem. So, McGinn’s criticism of panpsychism and panprotopsychism additionally works to distance his view from Russellian physicalism.

Finally, McGinn says of our understanding of the physical that “… [Locke] thought that our ideas of matter are quite sharply constrained by our perceptions and so concluded that the true science of matter is eternally beyond us … It looks today as if Locke was wrong about our capacity to fathom the nature of the physical world” (McGinn 1989, p. 352). Thus, for McGinn, the mystery of consciousness does not arise due to an inadequate account of the nature of the physical, and a theory like Russellian physicalism—which explains consciousness relative to certain fundamental physical properties—is doomed to fail. Rather, McGinn thinks the mystery lies at some sort of “psychophysical nexus”. What does this mean?

I believe McGinn’s view is the following: consciousness is identical to some garden-variety, scientifically scrutable property of the brain, but we cannot ever grasp how this could possibly be so. For
instance, he says there exists a scientific theory—inaccessible to cognitively limited human beings—which “… describes the link between consciousness and the brain in a way that is no more remarkable (or alarming) than the way we now describe the link between the liver and bile” (McGinn 1989, p. 361-62).

Russellian physicalism is a constructive account of consciousness which emphasizes our hitherto inadequate conception of the physical as responsible for the mind-body problem, which can be solved (in thought, if not in science) by expanding our notion of the physical. All the better for Russellian physicalism: McGinn’s view is too pessimistic about our chances of addressing the problem of consciousness. Mysterianism makes consciousness too mysterious—if McGinn were right, we could not conclusively say anything positive about subjective experience or its place in nature. Russellian physicalism, though pessimistic about any sort of reductive project in science that would explain consciousness in terms of scientifically scrutable properties, is not nearly as pessimistic as is full-blooded mysterianism.

That said, Russellian physicalism does bear a number of similarities to mysterianism. Both views acknowledge that consciousness is not scientifically scrutable, and both say that we are perceptually closed from access to the phenomenal properties of other mind-endowed beings. The biggest difference between the two views is that Russellian physicalism deems that expanding our conceptual repertoire to include (admittedly rather thin) concepts of categorical properties suffices to address the mind-body problem, and McGinn thinks that no such conceptual expansion could possibly be helpful.

While mysterianism rightly acknowledges the strangeness of consciousness, it goes too far. It says that consciousness cannot be understood by us in any way whatsoever. This is not at all satisfying: we would like to know something about what consciousness is, and why the nature of consciousness makes it so strange to us. Russellian physicalism plausibly does both tasks, and for this reason is preferable to the extreme mysterianism of McGinn.

3. The Phenomenal Concepts Strategy

Like mysterianism, the phenomenal concept strategy asserts that there is an unbridgeable epistemic gap between our scientific understanding of the world and an understanding of consciousness. However, unlike mysterianism, the phenomenal concept strategy explains where this epistemic gap comes from, and why it persists. Further, this strategy deflates the significance of the gap: though a satisfying
explanation of consciousness will forever elude us, we may nonetheless acquire an accurate scientific
explanation of the phenomenon. Stated in this way, the phenomenal concept strategy has an air of
paradox, but there is no contradiction in the view. Allow me to explain.

According to the phenomenal concept strategy, the reason we think that neural or other
scientifically scrutable properties cannot explain phenomenal properties is because we have phenomenal
concepts that are distinct from and incommensurable with our physical concepts. Though the contents of
these two sorts of concepts might be the same (e.g. they both are about neural properties, if type-identity
theory is true), the mode of presentation is different. Phenomenal concepts present phenomenal
properties in a way that does not allow us to see how the phenomenal properties could be identical to any
physical properties, regardless of the true nature of those phenomenal properties.

There are a variety of ways to cash out the nature of phenomenal concepts relative to different
modes of presentation. For instance, philosophers have suggested that phenomenal concepts are
indexical (Perry 2001), quotational (Papineau 2002), recognitional (Loar 1990, Carruthers 2003), or
inferential-role-defined (Hill and McLaughlin 1999). The differences between these various accounts won't
matter here—what matters is that all of these views say that phenomenal concepts present phenomenal
properties in a way that cannot be reconciled with how physical concepts present phenomenal properties.

The irreconcilability of phenomenal and physical concepts provides supporters of the phenomenal
concept strategy with a set of answers to typical anti-physicalist challenges. For instance, about
Jackson’s knowledge argument, they say that when Mary leaves the black-and-white room and first sees
red, she merely acquires a new phenomenal concept which presents the very same physical facts about
color perception that she was aware of from inside the room, but presents those facts in a different way.
Mary could infer knowledge of phenomenal properties from her physical knowledge while inside the room,
but was unable to acquire the phenomenal concept which presents the phenomenal properties in the way
they are presented when she actually sees red.

About Levine’s explanatory gap argument, the phenomenal concept strategist says that we do not
find any explanation of consciousness satisfactory because of the difference between our phenomenal
and physical concepts. The phenomenal concept presents facts about consciousness in one way, and the
physical concept presents those facts in a different way. There is an unbridgeable gap between the two
modes of presentation of the same thing, which leaves us unsatisfied with any physical account of consciousness. This is despite the availability to us of genuine scientific explanations of the physical properties that phenomenal properties are identical to.

Finally, about the conceivability argument, the phenomenal concept theorist says that though zombie worlds are conceivable, they are not possible. As with the other two arguments, the conceivability of zombie worlds follows from the difference between physical and phenomenal concepts: when we think of zombie worlds, we think of worlds which realize the contents of our physical concepts, and which do not realize the contents of our phenomenal concepts. This seems viable to us because of the radically different modes of presentation of those two sorts of concepts. As a matter of fact, though, such worlds are not genuinely possible, since phenomenal properties are identical to physical properties.

As I see it, the phenomenal concept strategy faces two serious objections that are not faced by Russellian physicalism. One is Chalmers’ (2007) “master argument” against the phenomenal concept strategy, which is easily responded to by the Russellian physicalist. The other issue is that the phenomenal concept strategy is ultimately an error theory: it says that Mary is wrong when she thinks she acquires new knowledge; we are wrong when we think that phenomenal properties are not explained by physical properties; and we are wrong when we come to believe that zombie worlds are possible. Such an error theory might be warranted if there were no other plausible physicalist explanation of consciousness, but this is not the case. Russellian physicalism provides such a non-error-theoretic account.

Chalmers’ master argument is the following:

1. Either we can conceive of zombie twins who possess the same phenomenal concepts as us, or we cannot

2. If zombie twins who possess the same phenomenal concepts as us are conceivable, then phenomenal concepts fail to explain our epistemic situation regarding consciousness

3. If zombie twins who possess the same phenomenal concepts as us are not conceivable, then phenomenal concepts are not physical

∴ 4. Either phenomenal concepts fail to explain our epistemic situation regarding consciousness, or they are not physical
“Zombie twins” in this argument refers to our zombie duplicates in a minimally physical zombie duplicate of our world (as I discussed in section one). The argument gets off the ground because phenomenal concept strategists accept the conceivability of zombie worlds, even though they deny the possibility of these worlds. The conclusion of the argument is a dilemma with two sharp points, neither of which are acceptable to the phenomenal concept theorist: phenomenal concepts are useless to explain the explanatory gap, or they are not physical. Let’s look at the details.

The first premise falls out of the law of excluded middle: either zombie twins can have the same phenomenal concepts as we do, or they cannot. Suppose zombies can possess the same phenomenal concepts as us, on the assumption that phenomenal concepts are entirely physical. Well, by stipulation, the zombies are in a different epistemic situation than we are: we are conscious, and cannot understand how physical properties explain phenomenal properties. Zombies are not conscious, so there is no genuine explanatory puzzle for them: they have no phenomenal properties which are in need of explaining. Yet, if we have exactly the same phenomenal concepts, we ought to be in exactly the same epistemic situation. No good!

On the other hand, suppose that zombies cannot possess the same phenomenal concepts as us. Yet how can this be if physicalism is true? Surely our minimal physical duplicates must possess the very same concepts as us! It seems that the only explanation for why we and our zombie twins could possibly possess different concepts is if the phenomenal concepts are at least partially constituted by something nonphysical. Yet that cannot be the case for a physicalist phenomenal concept strategist!

Various responses by phenomenal concept strategists have been offered, and the dialectic can get quite convoluted. For instance, Carruthers and Veillet’s (2007) response involves a sophisticated appeal to the difference between first-person-characterized and third-person-characterized phenomenal concepts (which they assert that Chalmers equivocates between). I am skeptical, but perhaps something like Carruthers and Veillet’s response might ultimately succeed as a viable response to the master argument. Regardless, Chalmers’ master argument presents a powerful prima facie objection to the phenomenal concept strategy, and this powerful argument is easily sidestepped by a Russellian physicalist.
Russellian physicalism, like the phenomenal concept strategy, also says that the mind-body problem exists because we have incommensurable physical and phenomenal concepts. The Russellian physicalist does not explain this incommensurability in terms of different modes of presentation, as the phenomenal concept strategist does; rather, phenomenal and physical concepts have different contents. Our robust physical concepts are of non-categorical properties, whereas our phenomenal concepts are of non-fundamental categorical properties. Since these properties are not identical, the Russellian physicalist can plausibly deny premise three of Chalmers’ argument: zombie twins who possess the same phenomenal concepts as us are not conceivable because our phenomenal concepts involve categorical properties, and the zombies’ “phenomenal” concepts do not—not because our phenomenal concepts involve anything nonphysical.

This is similar to the answer that nonphysicalists give. However, because Russellian physicalists expand the number of physically-acceptable ontological categories, the move is warranted for them without making phenomenal concepts at all nonphysical. As I discussed in section one, the Russellian physicalist must assert that we cannot conceive of zombie worlds that are true minimal physical duplicates of our world. We instead are conceiving of minimal non-categorical duplicates, which lack categorical properties. Zombies conceived of in this way do not have access to the same properties that we have access to in subjective experience.

Chalmers’ master argument presents a decisive reason to prefer Russellian physicalism over the phenomenal concept strategy. As I said, though, there may be a way for a phenomenal concept strategist to resist the noxious conclusion of Chalmers’ argument. Even if this is so, there is an additional reason Russellian physicalism is superior to the phenomenal concept strategy: Russellian physicalism is not an error theory, and the phenomenal concept strategy is.

To see why the phenomenal concept strategy is an error theory, suppose that phenomenal properties are identical to non-categorical neural properties. Can we scientifically explain these neural properties? I strongly suspect that we can: they are explained by lower-level biological, chemical, atomic and subatomic properties. Can we make correct inferences about these neural properties? Again, I see

---

38 Russellian physicalism is neutral between which specific account of phenomenal concepts is correct. However, probably something like Papineau’s (2002) quotationalism is the most appropriate candidate, since this allows phenomenal concepts to involve categorical phenomenal properties as constituents.
no reason why not: just as we can infer properties of complex non-mental systems from lower-level properties, we ought to be able to infer properties of the brain from the properties of the brain's parts. Under the assumption that phenomenal properties are neural properties, we are fully capable of inferring and explaining everything relevant about consciousness via a scientific method; nonetheless, we feel that these are incorrect inferences and insufficient explanations. The phenomenal concept strategy merely offers an explanation to account for our mistaken judgements about these explanations and inferences.

Under the phenomenal concept strategy, Mary does know everything there is to be known about the phenomenology of redness while still in the room! All that the phenomenal concept strategy adds is an explanation for why it seems to her that she learns about a new property when she leaves the room. It is an account of her error. Likewise, we think zombie worlds are possible, but we are wrong. The phenomenal concept strategy accounts for this error as well.

In my eyes, any such error theory is far too deflationary about the status of our phenomenal judgements. If the phenomenal concept strategy is right, then consciousness looks weird to us, but only because we have a silly bit of cognitive machinery that prevents us from seeing how phenomenal and physical concepts can both refer to a single thing. If our concepts were better behaved, then the mystery of consciousness would dissolve, since we would be able to recognize that phenomenal properties are identical to neural (or functional, or representational, or informational) properties. To put the same point more emphatically: the phenomenal concept strategy fails to account for Mary's tears of joy and wonder when she sees red for the first time.

39 A phenomenal concept strategist might object, and say that there is a sense in which Mary does learn a new fact when she sees red: she learns about a different mode of presentation. To borrow an example from David Papineau (forthcoming), this is similar to a case in which someone knows that Cary Grant is from Bristol, but doesn't know that Cary Grant is identical to Archie Leach. When she finds out that Archie Leach is from Bristol, it seems that she learns a new fact, even though the two propositions have identical wide-scope content. As I see it, this sort of response does not diminish my charge that the phenomenal concept strategy is an error theory: if the phenomenal concept strategy is right, then perhaps Mary may learn some sort of fact when she leaves the room, but not the fact that she thinks she learns. It does not seem to her that she learns about a new conceptual mode of presentation for an already-known fact—rather, she thinks she learns about a mental property that was previously unknown to her.

40 There is a puzzle about phenomenal judgements lurking here for Russellian physicalists: what accounts for the tears of wonder and joy of Mary's zombie twin? Though she has no feeling, and her emotions are thus ersatz wonder and ersatz joy, it nonetheless is the case that the non-categorical properties of zombie Mary are sufficient to account for her phenomenal judgements (see Chalmers
Compared to Russellian physicalism, the strangeness of consciousness is not sufficiently respected. Phenomenal concept strategists come closer than traditional reductive physicalists to properly accounting for consciousness’ strangeness, but are off the mark. Conversely, Russellian physicalists need not offer an error theory: we are not wrong when we judge that non-categorical properties fail to explain consciousness, or that we cannot infer phenomenal properties from non-categorical properties, or that zombies are (in some sense) possible. This provides a good reason to prefer Russellian physicalism over the phenomenal concept strategy: error theories should be avoided if they can be!

4. Conclusion
I have argued that, of all the available accounts of consciousness, Russellian physicalism is the most plausible contender. Nonphysicalism violates our Copernican attitudes, whereas traditional reductive physicalism fails to account for the strangeness of consciousness whatsoever. Mysterianism is too mysterious, and the phenomenal concept strategy is too deflationary. Only Russellian physicalism is in the sweet spot, providing a theoretical account of consciousness and its place in nature, without saying we are in error when we judge that scientific explanations fail to account for consciousness, and without making consciousness fundamental. Moreover, it offers plausible answers to all the standard anti-physicalist challenges, including an answer to Chalmers’ master argument against the phenomenal concept strategy.

However, even if it one is unconvinced that Russellian physicalism is the best view on consciousness of all those available, it should be acknowledged that Russellian physicalism nonetheless constitutes significant progress on the mind-body problem. This is because the standard arguments for nonphysicalism are equally arguments for Russellian physicalism. This means that there are no longer any good reasons to be a nonphysicalist: all such reasons are neutral between nonphysicalism and Russellian physicalism. I suspect that there are Copernican philosophers who would like to be physicalists, and not posit human-defining properties at the bottom level of nature, but have come to see physicalism as hopeless due to the arguments of Chalmers, Jackson and company. Now there is no

forthcoming). Why aren’t non-categorical properties also sufficient to account for non-zombie Mary’s phenomenal judgements? I don’t claim to have the solution to this puzzle, but Alter and Coleman’s (in preparation) work on mental causation under Russellian monism may help: Mary’s phenomenal judgements are caused by phenomenology, and zombie Mary’s judgements are caused by non-categorical properties only. Different ultimate causal explanations must be given for these two worlds.
barrier for philosophers of this stripe to stop worrying about nonphysicalism, and they can restrict philosophical debate to the relative merits of competing physicalist accounts of consciousness.

There is one caveat: though Russellian physicalism is a version of physicalism in the *via negativa* sense, it is not a version of physicalism in the theory-based sense. This is because Russellian physicalism requires positing categorical properties which are in principle outside the scope of any possible human science. This means that Russellian physicalism, which is a quite attractive view on the nature of consciousness, cannot be adopted by physicalists who tie ‘physical’ to physics. It thus gives up on what I referred to as "scientific imperialism"—the hope that all of nature is in principle discernible by human science. This is the cost of answering anti-physicalist arguments in a way that a nonphysicalist would find acceptable.

Physicalists who would like to use the view developed in this paper to defeat anti-physicalist arguments are thus constrained to *via negativa* physicalism, with the attendant existential consequences that I identified in the previous section of this dissertation. This is because Russellian physicalism makes mentality categorical and scientifically inscrutable, but not fundamental. As a result, robust free will is impossible, as well as a straightforward version of an immortal soul (though a quasi-soul is still possible). However, since Russellian physicalism is only explicitly concerned with human mentality, it is a bit weaker than full-blooded *via negativa* physicalism: for all we know, there is a God who is a fundamentally mental being, and there are fundamental normative properties.

This means that one *could* accept Russellian physicalism without being a complete *via negativa* physicalist. However, the very reasons that motivate adopting Russellian physicalism also constitute reasons to accept a complete version of *via negativa*: Copernicanism should lead us to reject not only fundamental mental properties, but fundamental normative properties as well. The existence of such fundamental normative properties puts morality—an especially human-seeming phenomenon—at the bottom level of nature. It also seems that the existence of a fundamentally mental God would violate our Copernican attitudes. Hence, a Russellian physicalist who is aware of the motivations for her view ought to embrace all of *via negativa* physicalism, which means giving up on the possibility of ultimate purpose, karma, and God (though, again, pseudo-karma and pseudo-God are possible).
V. Bibliography


Bernstein, S. (forthcoming). Could a middle level be the most fundamental?


Schwitzgebel, E. (2015). If materialism is true, the United States is probably conscious. Philosophical Studies 172 (7):1697-1721.


Stoljar, D. (2015). Russellian Monism or Nagelian Monism?


