Analysis in Mind

by

Andrew Botterell

B.A. (Honors), McGill University, 1990

Submitted to the Department of Linguistics and Philosophy in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the Massachusetts Institute of Technology

September 1998

© 1998 Andrew Botterell. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part.

Signature of Author: ________________________________
Department of Linguistics and Philosophy

Certified by: ________________________________________
Robert Stalnaker
Professor, Department of Linguistics and Philosophy
Thesis Supervisor

Accepted by: ________________________________________
Vann McGee
Professor, Department of Linguistics and Philosophy
Chair of the Committee on Graduate Students
Analysis in Mind

by

Andrew Botterell

Submitted to the Department of Linguistics and Philosophy in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Philosophy

Abstract

From the time of Descartes to about the 1960s, a certain epistemological idea dominated the philosophy of mind, namely the idea that theses about the relation between mind and body are, if true, a priori truths. Much of recent philosophy of mind is devoted to the question whether that idea is right. My research is largely an attempt to argue that some recent defenses of it are unsuccessful.

For example, Physicalism is the metaphysical thesis that every actual psychological event, property, or process is necessitated by some actual physical event, property, or process. Many philosophers believe that Physicalism is true. Until about the 1960s, those who believed it true typically believed that statements relating mind and body were a priori truths. Let us call this thesis A Priori Physicalism. Many philosophers nowadays believe, instead, that statements relating mind and body are only a posteriori truths. Let us call this thesis A Posteriori Physicalism. A number of philosophers have argued in recent years that A Posteriori Physicalism is unacceptable; on their view, Physicalists had better be A Priori Physicalists. My thesis examines the question whether that view is correct.

I begin with a discussion of two influential arguments for the conclusion that Physicalists must be A Priori Physicalists. Chapter 1 addresses itself to an argument for the conclusion that if physicalism is true, every referring psychological expression is coreferential a priori with some referring physical expression. This argument is commonly called the Property Dualism Argument against Physicalism. I argue that the Property Dualism Argument rests on an ambiguous premise: on one reading it begs the question against A Posteriori Physicalism, on the other reading the conclusion of the Property Dualism Argument does not follow.

Chapter 2 addresses itself to an argument of Frank Jackson’s for the conclusion that Physicalists must have an a priori story to tell about how the physical nature of the actual world makes true the psychological nature of the actual world. I distinguish two ways in which this claim might be understood,
and I argue that on neither way of understanding it does Jackson have a compelling argument for A Priori Physicalism.

Finally, in Chapter 3 I turn to a more general discussion of the relation between conceivability and possibility, and its bearing on the dispute between A Priori and A Posteriori Physicalists. I focus in particular on a recent argument of David Chalmers' from the conceivability of so-called zombies to the conclusion that A Posteriori Physicalism is false. I argue that this argument fails to provide compelling reasons for rejecting A Posteriori Physicalism. I argue, first, that it misconstrues the relation between conceivability and possibility, and second, that it fails to establish that zombies are conceivable in the relevant sense.

Thesis Supervisor: Robert Stalnaker
Title: Professor, Department of Linguistics and Philosophy
Acknowledgments

I cannot overemphasize the extent to which my thesis, and my thinking on philosophical matters both connected and unconnected with it, have been shaped by the patient criticism and advice offered to me by the members of my thesis committee: Alex Byrne, Robert Stalnaker, and Judith Jarvis Thomson. I consider myself extraordinarily fortunate to be able to count them as my teachers. I have learned an enormous amount from each of them—although not, I think, at a rate they would have chosen—and for this I am very grateful. In addition, I would like to thank Ned Block, both for reading and commenting on earlier versions of various chapters of this thesis, and for first getting me interested in the issues with which this thesis is concerned, and Noam Chomsky, for taking the time to discuss his views on language and the mind with me.

Many people have provided me with support, both intellectual and otherwise, over the past few years. Among those I would particularly like to thank are: Leonard Clapp, Mati Eklund, Judith Feldmann, Michael Glanzberg, Delia Graff, Timothy Hinton, David Hunter, Jennifer Noonan, Dean Pettit, Paul Pietroski, Lisa Sereno, Cara Spencer, Robert Stainton, Robert Streiffer, and, especially, Daniel Stoljar.

My graduate studies were funded in part by a grant from the Social Sciences and Humanities Research Council of Canada; I am grateful for their support.

Finally, I dedicate this to my parents, with love and no small degree of relief.
“It really is a nice theory. The only defect it has is probably common to all philosophical theories. It’s wrong.” (Saul Kripke, *Naming and Necessity*)

“[T]he history of philosophy is the history of smart people saying things that mostly aren’t true.” (Jerry Fodor)
Table of Contents

Abstract 3
Acknowledgments 5
Table of Contents 7

Chapter 1: The Property Dualism Argument against Physicalism 9

Chapter 2: Physicalism, Entailment, and Analysis 37

Chapter 3: Conceivability, Possibility, and the Very Idea of Zombies 75

References 111
Chapter One
The Property Dualism Argument against Physicalism

1. In its simplest form, physicalism in the philosophy of mind—hereafter *physicalism*—is the metaphysical thesis that the psychological nature of the actual world is wholly physical. So in particular, if there are psychological states, events, processes, or properties, then the physicalist says that they are identical with physical states, events, processes, or properties.¹ This metaphysical thesis about the psychological nature of the actual world is often supplemented with the further epistemological claim that it is possible to defend physicalism on a posteriori grounds alone. According to this version of physicalism—which I will call *a posteriori physicalism*—no a priori story needs to be told about the relation between psychological nature and physical nature in order for physicalism to be true. More precisely:

**A posteriori physicalism**

(i) For every referring psychological expression \( \psi \) there is a referring physical expression \( \phi \) such that \( \psi = \phi \) is true; and (ii) wherever \( \psi \) is a referring psychological expression and \( \phi \) is a referring physical expression and \( \psi = \phi \) is true, \( \psi = \phi \) is true only a posteriori.²

¹ For the purposes of this paper, I will concentrate on identity thesis versions of physicalism. What I say here about identity thesis versions of physicalism also goes over, with appropriate changes, to supervenience versions of physicalism. For arguments to this effect see Chapter 2. For similar arguments see Byrne (1998), Block and Stalnaker (1997), and Yablo (1998).

² For the remainder of this paper, use of the variable ‘\( \psi \)’ will indicate quantification over referring psychological expressions; use of the variable ‘\( \phi \)’ will indicate quantification over referring physical expressions.
A posteriori physicalism is an attractive thesis.\(^3\) Despite its attractiveness, however, a familiar argument alleges that a posteriori physicalism cannot be true. This argument was articulated most famously by J.J.C. Smart in his classic paper “Sensations and Brain Processes”. It has received renewed attention by Stephen White in his well-known paper “Curse of the Qualia”. This argument is sometimes called the \textit{Property Dualism Argument against Physicalism}, or the \textit{Property Dualism Argument} for short. My aim in this paper is to consider White’s version of the Property Dualism Argument and to argue that it fails.

2. Since this paper is concerned with physicalism, a criterion which would enable us to clearly distinguish psychological entities from physical entities might seem to be required. Despite the importance of such a criterion, however, I will leave the distinction between psychological entities and physical entities purposefully vague. There are two reasons for this. First, a rough, intuitive grasp of the distinctions is all that is required to state and understand the Property Dualism Argument. And second, it is notoriously difficult to make these distinctions more precise, and I have no very good suggestions about how this might be done. Some remarks, however, may prove helpful.

By a physical property I have in mind a property which can be had by paradigmatic non-conscious objects, or a property which is constituted by properties which can be had by paradigmatic non-conscious objects. Thus,

suppose the predicate ‘is F’ is part of a completed physical theory, where a completed physical is a theory which suffices in principle to explain the nature of all non-conscious objects.\textsuperscript{4} Then if $P$ is the property for which the predicate ‘is F’ stands, $P$ will be a physical property. I will call any expression which is part of a completed physical theory a physical expression.

What properties might count as physical properties? This is a difficult question to answer. Perhaps the best we can do is point to some uncontroversial examples of non-conscious objects and say that any properties which are needed to explain their natures will count as physical properties. This is obviously rough, and leaves certain important issues unresolved.\textsuperscript{5} What are we to say about the property of being a mountain, for instance: is it a physical property? It might be thought that it is not, since the predicate ‘is a mountain’ will likely not form part of a completed physical theory. But surely the property of being a mountain is a physical property if any property is. Here I think we can only say that since mountains are paradigmatic non-conscious objects, the predicate ‘is a mountain’ must be a physical predicate, and hence, that the property of being a mountain must count as a physical property.

Psychological properties next. By a psychological property I have in mind a property which can be had only by objects that sometimes think, or are sometimes conscious. Thus suppose ‘is F’ is a predicate that can by truly

\textsuperscript{4} See Block (1978) for a definition of ‘physical property’ along similar lines.

\textsuperscript{5} In particular, if panpsychism is true and paradigmatic non-conscious objects have psychological properties, then it will prove impossible to distinguish physical properties from non-physical properties in this manner. Panpsychism is not a plausible position, however, and there is very little reason to believe that it is true. After all, can it seriously be thought that rocks have (positive) psychological properties? It seems to me that this is not a possibility that we have to take seriously. For this reason, I will ignore the objection from panpsychism in what follows.
applied only to objects that sometimes think, or are sometimes conscious. Then if \( P \) is the property for which 'is \( F \)' stands, \( P \) will be a *psychological property*. Paradigmatic examples of such properties are intentional properties—such as the property of believing that pigs fly—and qualitative properties—such as the property of being in pain. I will call any predicate which stands for a psychological property a *psychological predicate*. Any expression which contains a psychological predicate I will call a *psychological expression*.\(^6\)

3. I turn now to Stephen White’s version of the Property Dualism Argument.

The Property Dualism Argument is designed to reduce to absurdity a physicalist position that identifies qualitative psychological states, such as Smith’s being in pain at \( t \), with physical states of the brain, and does so a posteriori. White argues as follows:

I am assuming, for simplicity, that a person’s qualitative state of pain at \( t \), say Smith’s, is identical with a physical state, say Smith’s brain state of type X at \( t \). Even if this is the case, however, not only do the sense of the expression ‘Smith’s pain at \( t \)’ and the sense of the expression ‘Smith’s brain state X at \( t \)’ differ, but the fact that they are coreferential

\(^6\) Given my definitions of ‘physical property’ and ‘psychological property’, it might seem that the thesis of a posteriori physicalism is trivially false. For consider some physical property \( P \), and suppose that ‘is \( F \)’ is the physical predicate which stands for \( P \). Then the expression ‘the \( x \) such that \( x \) is \( F \) or \( x \) is in pain’ will count as a physical expression—in light of the fact that it contains the physical predicate ‘is \( F \)’—and it will also count as a psychological expression—in light of the fact that it contains the psychological predicate ‘is in pain’. Moreover, since every expression is coreferential a priori with itself, and since any psychological expression can be turned into a physical expression by disjoining it with some physical expression, it might seem that every psychological expression is coreferential a priori with some physical expression. I allow that given this very loose notion of a physical expression, it is true that every psychological expression is coreferential a priori with some physical expression; what I deny is that this shows that the thesis of a posteriori physicalism is incoherent. This sort of objection will be dealt with in more detail below. (I am indebted here to Alex Byrne.)
cannot be established on a priori grounds. Thus there must be different properties of Smith’s pain (i.e., Smith’s brain state X) in virtue of which it is the referent of both terms... The general principle is that if two expressions refer to the same object and this fact cannot be established a priori, they do so in virtue of... different modes of presentation of that referent... The natural candidates for these modes of presentation are properties...

Since there is no physicalist description that one could plausibly suppose to be coreferential a priori with an expression like ‘Smith’s pain at t’, no physical property of a pain (i.e., a brain state of type X) could provide the route by which it was picked out by such an expression... (White 1986, 91-4)

A general argument is suggested by White’s remarks. Some terminology is required before we can turn to it, however.

It will prove useful in what follows to have a perspicuous way to talk about an object’s being a certain way, or equivalently, about an object’s having a certain property. I will therefore help myself to the notion of a state description, where a state description is an expression of the form ‘x’s being F’, and where a state description describes a state of affairs. More precisely, if P is the property for which the predicate ‘is F’ stands, then the state description ‘x’s being F’ will describe the state of affairs that consists in x’s having P. Any state description that describes a state of affairs that consists in an object’s having a physical property I will call a physical state description. Any state description that describes a state of affairs that consists in an object’s having a psychological property I will call a psychological state description.

With these terminological matters out of the way, we can proceed to the Property Dualism Argument. As noted, this argument is designed to reduce to absurdity the thesis of a posteriori physicalism:
A posteriori physicalism:

(i) For every referring psychological expression $\psi$ there is a referring physical expression $\phi$ such that $[\psi = \phi]$ is true; and (ii) wherever $\psi$ is a referring psychological expression and $\phi$ is a referring physical expression and $[\psi = \phi]$ is true, $[\psi = \phi]$ is true only a posteriori.

The core of the argument can be stated as follows:

(P1) For all expressions $\alpha$ and $\beta$, if $[\alpha = \beta]$ is true a posteriori then $\alpha$ and $\beta$ pick out their common referent in virtue of distinct properties of that referent.\(^7\)

The notion of an expression picking out its referent in virtue of its referent having a particular property is obscure, but let us leave it unanalyzed for the moment. I will turn to discussion of it presently.

For the purposes of our reductio, let us focus on some particular psychological state description, such as 'Smith’s being in pain at $t$'. Now, from the thesis of a posteriori physicalism it follows that

(C1) For every physical expression $\phi$, if 'Smith’s being in pain at $t = \phi$' is true, then 'Smith’s being in pain at $t = \phi$' is true only a posteriori.

---

\(^7\) There is a complication. The referring expressions I will be primarily concerned with are state descriptions, and I will sometimes say that the state description ‘$x$’s being $F$’ picks out its referent in virtue of its referent having some property $P$. This is strictly speaking incorrect. For the object $x$ is what has $P$, not the state of affairs that consists in $x$’s having $P$. What I should say is that ‘$x$’s being $F$’ picks out its referent in virtue of $P$, or in virtue of $x$’s having $P$, without claiming that the referent of ‘$x$’s being $F$’ has $P$. Since I don’t believe that anything depends on this issue, however, I will ignore it in what follows.
And from (P1) and (C1) it follows that

(C2) For every physical expression \( \phi \), ‘Smith’s being in pain at \( t' \) picks out its referent in virtue of a property that is distinct from the property in virtue of which \( \phi \) picks out its referent.

So far so good.

According to White, however, if ‘Smith’s being in pain at \( t' \) picks out its referent in virtue of a property that is distinct from the property in virtue of which any physical expression picks out its referent, then ‘Smith’s being in pain at \( t' \) picks out its referent in virtue of a non-physical property of its referent. For simplicity, let’s call the property in virtue of which ‘Smith’s being in pain at \( t' \) picks out its referent ‘Charlie’. Then White’s claim is that Charlie is a non-physical property.

In order to appreciate the force of this claim, however, we need to distinguish two ways in which a property might be non-physical. Call a property weakly non-physical if having that property does not entail being purely physical in nature. Call a property strongly non-physical if having that property does entail being non-physical in nature. Strongly non-physical properties are what might also be called irreducibly mental properties, i.e., mental properties which are distinct from all physical properties. This being the case, it is clear that White’s claim that Charlie is a non-physical property is best interpreted as the claim that Charlie is a strongly non-physical property.

To recapitulate, White’s claim is that if (C2) is true, then Charlie is a strongly non-physical property. Unfortunately, there is a gap in White’s reasoning: (C2) is not equivalent to, nor does it entail, the claim that Charlie is a strongly non-physical property. For even if Charlie is not identical with a
property in virtue of which any physical expression picks out its referent, this is consistent with Charlie’s being a physical property so long as no physical expression picks out its referent in virtue of its referent having Charlie.

In short, an additional premise is needed to get from (C2) to the conclusion that Charlie is a strongly non-physical property. The premise suggested by White’s discussion is:

(P2) If a psychological expression $\psi$ picks out its referent in virtue of its referent having a physical property $F$, then there is a physical expression $\phi$ that also picks out its referent in virtue of its referent having $F$.

If (P2) is conjoined with (P1), it follows that a posteriori physicalism is false.

To see why, suppose that ‘Smith’s being in pain at $t$’ picks out its referent in virtue of its referent having Charlie, and suppose further that Charlie is a physical property. Then by (P2) there will be a physical expression $\phi$ which also picks out its referent in virtue of its referent having Charlie. But if ‘Smith’s being in pain at $t$’ and $\phi$ pick out their referent in virtue of the same property of that referent, then ‘Smith’s being in pain at $t$’ and $\phi$ will be coreferential a priori, falsifying clause (ii) of the thesis of a posteriori physicalism. So on the assumption that Charlie is a physical property, a posteriori physicalism is false.

On the other hand, suppose that Charlie is a strongly non-physical property. Then there will be no physical expression with which ‘Charlie’ is coreferential, falsifying clause (i) of the thesis of a posteriori physicalism. So on the assumption that Charlie is a strongly non-physical property a posteriori physicalism is false.
In short, from (P1) and (P2) it follows that

\[(C)\] For all psychological expressions \(\psi\), either \(\psi\) picks out its referent in virtue of its referent having a strongly non-physical property, or there is a physical expression with which \(\psi\) is coreferential a priori.

is true. And (C) clearly entails that a posteriori physicalism is false.

The Property Dualism Argument is designed to show that a posteriori physicalism is incoherent. For according to the Property Dualism Argument, the conjunction of (P1) and (P2) entails that anyone who holds that the relation between psychological entities and physical entities is knowable only a posteriori cannot also hold that every psychological entity is a physical entity.

4. Thus far I have been primarily concerned with the structure of the Property Dualism Argument. I now want to focus on its soundness. To do this, however, I need to say something about the as yet unexplained notion of an expression picking out its referent in virtue of its referent having a particular property. In order to facilitate discussion I’ll say that if an expression \(a\) refers to its referent in virtue of its referent having a particular property \(P\) then \(a\) introduces \(P\). White says very little by way of explanation

---

8 I am assuming, as does White, that at least some psychological expressions refer. It is worth noting, however, that rejecting this assumption gives rise to another response to the Property Dualism Argument. For if psychological expressions do not refer, then \textit{a fortiori} they do not pick out their referents in virtue of their referents having certain properties, and the conclusion that some properties are strongly non-physical properties is blocked. For present purposes, however, I will ignore this eliminativist option.

9 This terminology is due to Brian Loar (1990). In addition, I will assume that while the relation of introducing is—at least in principle—many-one, it is not one-many. In other words, I will assume that while distinct expressions might introduce the same property, a given
of this notion. Still, given White's references to modes of presentations, two natural ways to understand the notion of an expression introducing a property suggest themselves. In brief, 'introduce' can be interpreted in a broadly Fregean or in a broadly Kripkean manner.

According to the broadly Fregean interpretation of 'introduce' I have in mind—the *Fregean interpretation* for short—an expression \( a \) is said to introduce a property \( P \) just in case \( P \) is the sense of \( a \). Thus, suppose 'water' is synonymous with the definite description 'the stuff that falls from the sky in the form of rain, fills the oceans and lakes, and is a colorless, odorless liquid'. Then on the reasonable assumption that the predicate 'is the stuff that falls from the sky in the form of rain, fills the oceans and lakes, and is a colorless, odorless liquid' stands for the property of being the stuff that falls from the sky in the form of rain, fills the oceans and lakes, and is a colorless, odorless liquid, then that property will be the sense of 'water', and so will be introduced by 'water'. More generally, according to the Fregean interpretation of 'introduce', if an expression \( a \) is synonymous with a description \( D \), then \( a \) will have as its sense the property for which the predicate 'is \( D \)' stands.

I turn now to the broadly Kripkean interpretation of 'introduce'. Kripke (1971, 1980) defends a semantic theory according to which proper names and natural kind terms pick out their referents via reference-fixing properties, but without the mediation of a Fregean sense. This suggests another way to make sense of the claim that expressions introduce properties. According to the broadly Kripkean interpretation of 'introduce' I have in mind—the *Kripkean interpretation* for short—an expression \( a \) is said to introduce a property \( P \) just

expression will introduce only a single property, although the property thus introduced might be a complex conjunctive one.
in case $P$ fixes the reference of $a$. Thus, suppose that the reference of ‘Bill Clinton’ is fixed by the property of being the President of the United States.\textsuperscript{10} If so, then ‘Bill Clinton’ will introduce the property of being the President of the United States. Importantly, since according to Kripke fixing the reference of an expression is not the same as giving the meaning of that expression, a property $P$ can fix the reference of an expression $a$ even if $P$ is associated with $a$ only a posteriori.\textsuperscript{11} In consequence, the Kripkean interpretation of ‘introduce’ differs from the Fregean interpretation discussed above.

5. So much by way of terminology; let us return to the Property Dualism Argument. To recapitulate, I have argued that White’s reductio of the thesis of a posteriori physicalism should be interpreted as consisting of two premises,

(P1) For all expressions $a$ and $\beta$, if $\alpha = \beta$ is true a posteriori then $a$ and $\beta$ pick out their common referent in virtue of distinct properties of that referent.

and

\textsuperscript{10} Equivalently: suppose that the reference of ‘Bill Clinton’ is fixed by the description ‘the President of the United States’. Then on the plausible assumption that the predicate ‘is the President of the United States’ stands for the property of being the President of the United States, the property of being the President of the United States will fix the reference of ‘Bill Clinton’. In short, I assume that the claim that the reference of an expression $a$ is fixed by a definite description $D$ is equivalent to the claim that the reference of $a$ is fixed by the property for which the predicate ‘is $D$’ stands.

\textsuperscript{11} Kripke also argues that reference-fixing properties can sometimes be associated with expressions a priori. I will return to this issue below.
(P2) If a psychological expression ψ picks out its referent in virtue of its referent having a physical property P, then there is a physical expression φ such that φ also picks out its referent in virtue of its referent having P.

and a conclusion

(C) For all psychological expressions ψ, either ψ picks out its referent in virtue of its referent having a strongly non-physical property, or there is a physical expression with which ψ is coreferential a priori.

This suggests three ways to resist the Property Dualism Argument: argue that at least one of (P1) or (P2) is false, or argue that (C) does not follow from the conjunction of (P1) and (P2) with the thesis of a posteriori physicalism. My argument will be simple. We have seen that the crucial notion of an expression introducing a property admits of two different interpretations, a broadly Fregean interpretation and a broadly Kripkean one. This results in two different versions of the Property Dualism Argument. I will argue that if ‘introduce’ is interpreted in a broadly Fregean manner then, first, there is no reason to believe (P1), and second, (P2) begs the question against a posteriori physicalism. I will then argue that if ‘introduce’ is interpreted in a broadly Kripkean manner, there is no reason to believe (P1). On neither interpretation is the Property Dualism Argument sound.

6. Let us begin by asking how the Property Dualism Argument fares if ‘introduce’ is interpreted in a broadly Fregean manner. If ‘introduce’ is
interpreted in a broadly Fregean manner we get the following two interpretations of (P1) and (P2):

(P1-Fregean)

For all expressions $a$ and $\beta$, if $\langle a = \beta \rangle$ is true only a posteriori then for some predicate $F$, $a$ is synonymous with $\text{the } F'$, and for some predicate $G$, $\beta$ is synonymous with $\text{the } G'$, and $\text{the property of being } F = \text{the property of being } G'$ is false.

(P2-Fregean)

If a psychological expression $\psi$ picks out its referent in virtue of $\psi$ having a physical property $P$ as its sense, then there is a physical expression $\phi$ such that $\phi$ also picks out its referent in virtue of $\phi$ having $P$ as its sense.

An immediate problem with the Fregean version of the Property Dualism Argument is that there is good reason to suppose that (P1-Fregean) is false. For as Kripke (1971, 1980) has persuasively argued, it is plausible to suppose that proper names and natural kind terms pick out their referents without such expressions being synonymous with any description whatsoever. This suggests that (P1-Fregean) is misconceived as a general semantic principle. However, some will no doubt continue to find the Fregean interpretation attractive. So let us grant (P1-Fregean) for the sake of argument. Is there reason to suppose that (P2-Fregean) is also true and hence, that the Fregean version of the Property Dualism Argument is sound? I will argue that there is not.
According to (P2-Fregean), if a psychological expression \( \psi \) picks out its referent in virtue of \( \psi \) having a physical property \( P \) as its sense, then there will be a physical expression \( \phi \) such that \( \phi \) also picks out its referent in virtue of \( \phi \) having \( P \) as its sense. It is worth noting that (P2-Fregean) is not without advocates. For example, Brian Loar appeals to something like (P2-Fregean) when he argues that if a physical property \( P \) were the sense of some psychological expression \( \psi \) "there would be an a priori connection between that [psychological] term [\( \psi \)] and some physical term, viz., one that more explicitly expresses that sense [i.e., \( P \)]." (Loar 1990, 84, n. 5) Loar is not arguing for (P2-Fregean) here, but is merely stating what he takes to follow from the adoption of something like (P1-Fregean). However, since Loar does not provide an explicit argument for his claim, it stands to reason that he thinks that it is simply obvious that a physicalist who endorses something like (P1-Fregean) is committed to thinking that if \( P \) is a physical property which is the sense of some psychological expression \( \psi \), then there is some physical expression \( \phi \) which also has \( P \) as its sense. But is this obvious?

It might be thought that (P2-Fregean) is clearly a principle that physicalists are committed to in virtue of being physicalists. For physicalists are committed to holding that a complete physical description of the nature of the actual world is a complete description of the nature of the actual world simpliciter. And at a minimum, it seems that such a description must include a specification of the relations that hold between physical properties and physical expressions. Moreover, since one way for a physical property \( P \) to be related to a physical expression \( \phi \) is for \( P \) to be the sense of \( \phi \) it might seem to follow from this that physicalists are committed to holding that a complete physical description of the actual world will say, for any physical property \( P \), which physical expression introduces \( P \). Hence, it would appear that there is
reason for thinking that physicalists are committed to holding that if $P$ is a physical property which is the sense of a psychological expression $\phi$ then $P$ is also the sense of some physical expression $\psi$.

I think it is clear, however, that this argument is mistaken. For even if it is granted that a physical theory must say, for any physical property $P$, whether $P$ is the sense of a physical expression, it does not follow from this that if a physical property $P$ is the sense of a psychological expression $\phi$, then $P$ is also the sense of some physical expression $\psi$. That is, even if being the sense of is a relation that holds between physical properties and physical expressions, and even if a completed physical theory must describe such relations, it does not follow that for every physical property $P$ there is some physical expression which has $P$ as its sense.

Perhaps it will be argued that (P2-Fregean) follows from physicalism because every physical property is the sense of some physical expression or other. Thus suppose $P$ is a physical property, and suppose $\text{is } F$ is the physical predicate which stands for $P$. Then given the predicate $\text{is } F$ there would seem to be no bar to our forming the expression $\text{the } F$. And on the assumption that an expression of the form $\text{the } F$ has as its sense the property for which the predicate $\text{is } F$ stands, every physical property will be the sense of some physical expression or other.\footnote{Another way of making this point would be to employ a modified version of what David Lewis (1986) has called a Lagadonian language. A Lagadonian language is a language in which every object in the domain of discourse names itself. The modified Lagadonian language I have in mind is a language in which every physical property is its own sense. If this modified Lagadonian proposal were adopted it would then be true that every physical property is the sense of some physical expression or other, since it would be true that every physical property is its sense. Whatever the merits of this modified Lagadonian proposal, however, it does not seem to be something that a posteriori physicalists are committed to in virtue of being a posteriori physicalists, and so can safely be set aside.} This is an important consideration, and I will
return to discussion of it below. For the moment, however, two things are worth noting.

First, the claim is that the sense of a definite description ‘the \( F \)’ is the property for which the predicate ‘is \( F \)’ stands. It would seem to follow from this, however, that if two predicates stand for the same property that it is a priori that they stand for the same property. But there are reasons for supposing that this is false: there are instances where two predicates stand for the same property only a posteriori. For example, while it is plausible to suppose that the predicate ‘is water’ and the predicate ‘is H\(_2\)O’ stand for the same property, it is arguable that ‘the property of being water = the property of being H\(_2\)O’—and hence, ‘the \( x \) such that \( x \) is water = the \( x \) such that \( x \) is H\(_2\)O’—cannot be known to be true a priori.

Second, it is natural to think that not just any property can be the sense of an expression. In particular, it is natural to think that any property which is the sense of an expression must be conceptually linked to the expression whose sense it is, and must also be the kind of property which can determine reference. However, since it is far from obvious that every physical property is both conceptually linked to some physical expression or other and is capable of determining reference in the manner required by (P\(_2\)-Fregean), this casts doubt on the above claim.

I realize that these remarks are inconclusive, and I don’t pretend to have a knock-down refutation of the claim that every physical property is the sense of some physical expression. Still, as I have indicated, I think that there are reasons to be skeptical about this claim.

7. It is arguable, then, that (P\(_2\)-Fregean) does not obviously follow from (P\(_1\)-Fregean), nor does it follow from the metaphysical thesis of physicalism.
But is (P2-Fregean) true nonetheless? It might be thought that the following considerations show that it is.

Suppose $P$ is a physical property which is the sense of some psychological expression $\psi$. Then either $P$ is the sense of an expression of some completed physical theory $L$—that is, either $P$ is the sense of some physical expression $\phi$—or it is not. If $P$ is the sense of an expression of $L$, then (P2-Fregean) is true. If $P$ is not the sense of an expression of $L$ then, given the existence of a physical theory like $L$, there is no obvious reason why there couldn’t be another physical theory, $L^*$, which is just like $L$ except for containing an additional physical expression that has $P$ as its sense. To see that the existence of a theory like $L^*$ is not in doubt, we need only consider the following. Take $L$. Add to it a new physical expression $\phi^*$. Call the new physical theory $L^*$. Stipulate that $\phi^*$ has $P$ as its sense. Then, since by hypothesis $P$ is the sense of a psychological expression and since by stipulation $P$ is the sense of a physical expression, namely $\phi^*$, it would appear that if $P$ is the sense of a psychological expression then it must also be the case that $P$ is the sense of a physical expression, in which case (P2-Fregean) is true.

While there are a number of things that might be said about the above line of argument, I think it is clear that a posteriori physicalists should not be persuaded by it. Instead, a posteriori physicalists should deny that $L^*$ is relevant to the truth or falsity of (P2-Fregean). This might seem odd. For it might be thought that if $L$ is an acceptable physical theory then there is no obvious reason for thinking that $L^*$ is not an acceptable physical theory. After

---

13 It might be objected that, since there must be some constraints on what counts as an acceptable physical expression, the proponent of the Property Dualism Argument cannot simply stipulate that $\phi^*$ introduces $P$ as a sense. I am sympathetic to this objection, and will discuss it in detail below.
all, $L^*$ contains only physical expressions which refer to physical entities by having physical properties as senses. And $L^*$ differs from $L$ only in containing an additional expression which has as its a physical property which was not the sense of any expression of $L$. But it is hard to see why this fact could entail that $L^*$ is not an acceptable physical theory. And if $L^*$ is an acceptable physical theory, then surely it is relevant to the truth or falsity of (P2-Fregean).

Nonetheless, I think that a posteriori physicalists have good reason to object to the relevance of $L^*$. First, note that $L$ is by hypothesis a completed physical theory. Consequently, it is unclear that a proponent of The Property Dualism Argument can appeal to a physical theory like $L^*$ without begging the question against a posteriori physicalism. For if $L$ is a completed physical theory, and if adding another expression to $L$ does not affect its explanatory power, then the addition of such an expression to $L$ would appear to be unmotivated, by which I simply mean that no additional expressive power would be gained by such an addition. And if adding another expression to $L$ does affect its explanatory power, then it would seem to follow that $L$ wasn’t complete to begin with. But then it would appear that a completed physical theory $L$ must be such that, if a physical property $P$ is the sense of a psychological expression $\psi$, then $L$ must contain an expression which has $P$ as its sense. But such a definition of ‘completed physical theory’ is equivalent to (P2-Fregean), since it simply assumes that there will be a physical expression which is coreferential a priori with $\psi$. And since the truth of (P2-Fregean) is what is at issue, it would beg the question against an a posteriori physicalist to assume a definition of ‘completed physical theory’ which takes the truth of (P2-Fregean) for granted.

Second, recall that the conclusion of the Property Dualism Argument is that unless there are physical expressions with which psychological
expressions are coreferential a priori, physicalists will be forced to acknowledge the existence of strongly non-physical properties. The above argument, however, only establishes that if a physical property \( P \) is the sense of a psychological expression \( \psi \) there \textit{could be} a physical expression which also has \( P \) as its sense. However, on one way of understanding this claim, it is irrelevant to the question whether a posteriori physicalism is true; and on the other way of understanding this claim, it trivializes the debate between a posteriori and a priori physicalists.

The first way of understanding the claim is that for all the a posteriori physicalist has argued there \textit{could be} physical expressions with which psychological expressions are coreferential a priori. However, physicalists can surely agree with this and still deny that the proponent of the Property Dualism Argument has thereby established the incoherence of a posteriori physicalism. \textit{Of course} there could be physical expressions with which psychological expressions are coreferential a priori; the question, however, is whether, if physicalism is true, there \textit{must be} physical expressions with which psychological expressions are coreferential a priori. So on this way of understanding the claim it would not appear to present any problem for a posteriori physicalism.

The other way of understanding the claim is that whenever a physical property \( P \) is the sense of a psychological expression but is not the sense of any physical expression, we can always create a new physical expression and stipulate that the newly created physical expression has \( P \) as its sense. However, if this is how the claim is to be understood, then it seems clear that the Property Dualism Argument is rendered completely uninteresting. For while it is true that new expressions can always be introduced into a physical theory, and can be stipulated to be synonymous with existing psychological
expressions, it is hard to see any threat to the coherence of a posteriori physicalism here.

It seems to me that physicalists who find this argument for (P2-Fregean) compelling must think so in part because they are tacitly appealing to something similar to what Alonzo Church has called the ‘Principle of Tolerance’. According to Church, the Principle of Tolerance asserts that “everyone is at liberty to build his own form of language as he will.” (Church 1954, 160 [italics in original]) So in particular, according to the Principle of Tolerance physicalists are at liberty to build their own form of physical language. Moreover, since one way to build a physical language is to build a language in which for every psychological state description \( \psi \), say, there is some physical state description \( \phi \) with which \( \psi \) is coreferential a priori, it might be thought that something like the Principle of Tolerance supports (P2-Fregean), and counts against a posteriori physicalism.

Now it does seems to me that if something like Church’s Principle of Tolerance is adopted then the truth of (P2-Fregean) can be established. But it also seems to me that establishing the truth of (P2-Fregean) in this manner makes it unclear what is at issue between a posteriori and a priori physicalists. A posteriori physicalism, recall, is the thesis that there is no a priori or conceptual connection between physical nature and psychological nature. But no a posteriori physicalist should deny Church’s Principle of Tolerance understood as the claim that, for any expression \( \alpha \) we can always create another expression \( \beta \) with which \( \alpha \) is coreferential a priori. What the a posteriori physicalist should deny is that this possibility shows that a posteriori physicalism is incoherent, any more than it shows that a posteriori chemistry is incoherent. By appealing to something like the Principle of Tolerance in order to establish the truth of (P2-Fregean) we run the risk of
misrepresenting what is really at issue between those physicalists who think that physicalism can be defended on a posteriori grounds only and those physicalists who think that physicalism requires that statements relating mind and body are a priori truths.

Here, then, we come to the fundamental problem with (P2-Fregean), namely that (P2-Fregean) is simply the denial of the thesis of a posteriori physicalism, and so begs the question against a posteriori physicalism. For if (P2-Fregean) is true then every psychological expression will have the same sense as some physical expression, and so every psychological expression will be coreferential a priori with some physical expression. But this is precisely what the thesis of a priori physicalism claims. In short, if the Fregean interpretation of ‘introduce’ is adopted, then (P2-Fregean) is equivalent to the denial of the thesis of a posteriori physicalism, and the Fregean version of The Property Dualism Argument begs the question against a posteriori physicalism.

8. It seems to me, then, that there are compelling reasons to reject the Fregean version of the Property Dualism Argument. First, there is reason to think that (P1-Fregean) is false. And second, there is reason to think that (P2-Fregean) begs the question against a posteriori physicalism. So let us instead consider whether the Property Dualism Argument fares any better if ‘introduce’ is interpreted in a broadly Kripkean manner.

If ‘introduce’ is interpreted in a broadly Kripkean manner, we get the following interpretations of (P1) and (P2):
(P1-Kripkean)

For all expressions \( \alpha \) and \( \beta \), if \( \alpha = \beta \) is true a posteriori then \( \alpha \) and \( \beta \) pick out their common referent in virtue of distinct reference-fixing properties \( F \) and \( G \), \( \alpha \) picking out its referent in virtue of \( F \) and \( \beta \) picking out its referent in virtue of \( G \).

(P2-Kripkean)

If a psychological expression \( \psi \) picks out its referent in virtue of a reference-fixing physical property \( P \), then there is a physical expression \( \phi \) such that \( \phi \) also picks out its referent in virtue of \( P \).

I propose to grant (P2-Kripkean) for the sake of argument. So let us turn to consideration of (P1-Kripkean) instead.

According to (P1-Kripkean) for all expressions \( \alpha \) and \( \beta \), if \( \alpha = \beta \) is true a posteriori then \( \alpha \) and \( \beta \) pick out their common referent in virtue of distinct reference-fixing properties \( F \) and \( G \), \( \alpha \) picking out its referent in virtue of \( F \) and \( \beta \) picking out its referent in virtue of \( G \). By contraposition we have that for all expressions \( \alpha \) and \( \beta \), if it is not the case that \( \alpha \) and \( \beta \) pick out their common referent in virtue of distinct properties of that referent, then it is not the case that \( \alpha = \beta \) is true a posteriori. Since we are assuming that \( \alpha = \beta \) is true—i.e., since we are assuming that \( \alpha \) and \( \beta \) are coreferential—the contrapositive of (P1-Kripkean) can be more perspicuously stated as the claim that if \( \alpha \) and \( \beta \) pick out their common referent in virtue of the same property of that referent, then \( \alpha = \beta \) is true a priori. What I wish to argue is that this

---

\[ \text{\textsuperscript{14}} \] I am therefore assuming that if an identity statement of the form \( \alpha = \beta \) is true, but not true a posteriori, then \( \alpha = \beta \) is true a priori.
claim is false if 'introduce' is interpreted in a broadly Kripkean manner. This will suffice to show that (P1-Kripkean) is false, and hence, that the Kripkean version of the Property Dualism Argument is unsound.

In general, it is a mistake to conclude from the fact that the reference of two expressions is fixed by the same reference-fixing property that those two expressions are coreferential a priori. To illustrate this point, consider a true a posteriori identity statement involving two definite descriptions 'the x such that $Fx'$ and 'the x such that $Gx'$ where the predicates is 'is $F'$ and is 'is $G'$ stand for the same property. Then according to the Kripkean interpretation of 'introduce' 'the x such that $Fx'$ will introduce as a reference-fixing property the property for which the predicate 'is $F'$ stands, and 'the x such that $Gx'$ will introduce as a reference-fixing property the property for which the predicate 'is $G'$ stands. But by hypothesis the property for which the predicate 'is $F'$ stands is the same property as the property for which the predicate 'is $G'$ stands. Thus the definite descriptions 'the x such that $Fx'$ and 'the x such that $Gx'$, although coreferential only a posteriori, introduce the same reference-fixing property, and so constitute a counter-example to the claim that if two expressions introduce the same reference-fixing property then those two expressions must be coreferential a priori.

How does this apply to the particular case of the psychological? Since we are granting (P2-Kripkean) for the sake of argument, we know that if the reference of a psychological state description such as 'Smith's being in pain at $t'$ is fixed by a physical property $P$ there will be a physical expression, such as 'Smith's being $F'$—where the physical predicate 'is $F'$ stands for $P$—whose reference is also fixed by $P$. So we know that if the reference of 'Smith's being in pain at $t'$ is fixed by a physical property $P$ there will be some physical expression $\phi$ whose reference is also fixed by $P$. Does it follow from this that
Smith’s being in pain at t’ is coreferential a priori with an expression like ‘Smith’s being F’? I do not see that it does.

To see why, we need only suppose—with the a posteriori physicalist—that ‘the property of being in pain’ and ‘the property of being F’—where ‘is F’ stands for a physical property—are coreferential only a posteriori. If this is the case, then although ‘Smith’s being in pain at t’ and ‘Smith’s being F’ will introduce the same reference-fixing property, they will not be coreferential a priori. In consequence, there is no apparent problem with claiming that a physical property P is introduced as a reference-fixing property by a psychological expression ψ, that P is also introduced as a reference-fixing property by a physical expression φ, but that ψ and φ are coreferential only a posteriori. In a way this should come as no surprise. For as noted above, the core of the Kripkean interpretation is that there need be no conceptual connection between an expression α and the property introduced by α as a reference-fixing property. And once the lack of such a conceptual connection is acknowledged, the inference from the claim that two expressions introduce the same reference-fixing property to the conclusion that those expressions must be coreferential a priori is blocked.

9. To this it might be objected that any interpretation of ‘introduce’ must entail that if two expressions α and β introduce the same property then α and β are coreferential a priori. The important point for our purposes, however, is that if this is indeed a constraint on any interpretation of ‘introduce’ then we have reason to think that the original argument in favor of (P2-Kripkean) is mistaken. There it was argued that if a physical property P fixes the reference of a psychological expression ψ, then because we can always form a physical state description involving the predicate that stands for P, there will be a
physical expression whose reference is also fixed by $P$. But consider two state 
descriptions, "Smith’s being in pain at $t'$" and "Smith’s being $F'$", which are by 
hypothesis coreferential a posteriori, but where the predicates ‘is in pain’ and 
‘is $F'$’ stand for the same physical property. By hypothesis ‘Smith’s being in 
pain at $t'$ and ‘Smith’s being $F'$’ are coreferential only a posteriori. But 
according to the constraint we are now considering, since the predicates ‘is in 
pain’ and ‘is $F'$’ stand for the same physical property, ‘Smith’s being in pain at 
$t'$ and ‘Smith’s being $F'$’ will introduce the same property, and so must be 
coreferential a priori. Contradiction.

This suggests one of two things. Either we must reject the idea that if 
two expressions $\alpha$ and $\beta$ introduce the same property then $\alpha$ and $\beta$ must be 
coreferential a priori, or we must reconsider our original reasons for finding 
(P2-Kripkean) plausible to begin with. If the former, then we no longer have a 
reason for supposing that (P1-Kripkean) is true. If the latter, then we no 
longer have an argument for (P2-Kripkean), and talk about expressions 
introducing properties becomes more mysterious than it already is. In either 
event, the Kripkean version of the Property Dualism Argument collapses.

Clearly, what is at issue is whether, on the Kripkean interpretation of 
‘introduce’, the reference-fixing properties which are associated with 
expressions are associated with those expressions a priori or only a posteriori. 
The preceding discussion has proceeded under the assumption that reference-
fixing properties are associated with expressions only a posteriori. But it 
might be objected that this ignores Kripke’s treatment of the contingent a 
priori, and Gareth Evans’ discussion of so-called ‘descriptive names’.

According to Kripke, “one should bear in mind the contrast between 
the a priori but perhaps contingent properties carried with a term, given by 
the way its reference was fixed, and the analytic (and hence necessary)
properties a term may carry, given by its meaning.” (Kripke 1980, 135). So, for example, Kripke holds that it is a priori that pain is what gives rise to pain-sensations, and thus that it is a priori that ‘pain’ has associated with it the property of giving rise to pain sensations.

Evans’ discussion focuses on the possibility of descriptive names. Consider his example of ‘Julius’. ‘Julius’, let us suppose, was introduced into our language as an abbreviation for the definite description ‘the actual inventor of the zip’. Then since the reference of ‘Julius’ is fixed a priori by the description ‘the actual inventor of the zip’, anyone who understands ‘Julius’ knows a priori that if anyone invented the zip, then Julius invented the zip. Evans calls such names ‘descriptive names’.

This suggests a view according to which psychological state descriptions such as ‘Smith’s being in pain at t’ are descriptive names. For such expressions pick out physical states—in this case a state of Smith’s brain—and it might be argued that they do so by having reference-fixing properties associated with them a priori. Moreover, it might be thought that if a physical property $P$ is associated as a reference-fixing property with a psychological expression $\psi$ a priori, then the anti-physicalist conclusions of the Property Dualism Argument follow.

Although the issues here are complex, I think to the contrary that this worry is misguided. Let us grant that if a physical property $P$ is associated as a reference-fixing property with a psychological expression $\psi$, then $P$ is either associated with $\psi$ a priori or only a posteriori. Still, if $P$ is associated with $\psi$ only a posteriori, then for the reasons just given there is reason to suppose that (P1-Kripkean) is false. On the other hand, if $P$ is associated with $\psi$ a priori—if, that is, there is a conceptual connection between $\psi$ and the property it introduces—then the Kripkean version of the Property Dualism Argument
collapses into the Fregean version, and criticisms similar to those brought against the Fregean version are available. In neither case does the Kripkean version of the Property Dualism Argument present a threat to a posteriori physicalism.

10. In summary, I have argued that the Property Dualism Argument provides no reason for physicalists to suppose that physicalism must take an a priori form. I distinguished two ways in which the crucial notion of an expression introducing a property might be interpreted. I first argued that if 'introduce' is interpreted in a broadly Fregean manner, then, first, there is good reason to suppose that (P1) is false, and second, that (P2) begs the question against a posteriori physicalism. I then argued that if 'introduce' is interpreted in a broadly Kripkean manner, (P1) is false. I therefore concluded that the Property Dualism Argument is unsound.

This observation is, I think, important. For many philosophers have employed versions of the Property Dualism Argument to argue that a priori analyses of psychological expressions are required if physicalism is to be true. I think this is a mistake. This is not to say that there might not be other reasons for thinking that a posteriori physicalism if false. If I am right, however, the conclusion of the present paper should lead those philosophers who have been convinced by the Property Dualism Argument to rethink their reasons for holding physicalism in an a priori, or conceptual, form.15

---

15 I would like to thank Ned Block, Alex Byrne, Noam Chomsky, Lenny Clapp, Ned Hall, Robert Stalnaker, Daniel Stoljar, and Stephen White for helpful discussions of these issues. I would especially like to thank Judith Jarvis Thomson for her patient advice and encouragement.
Chapter Two
Physicalism, Entailment, and Analysis

1. In several recent papers Frank Jackson has argued that philosophers who hold certain metaphysical views must also hold certain epistemological views. More precisely, Jackson has argued that physicalists in the philosophy of mind—hereafter physicalists—must have “an a priori story to tell about how the physical way the actual world is makes true the psychological way the actual world is.” (Jackson 1994, 40) I will have more to say about this thesis of Jackson’s, but for the moment let us simply call the claim that physicalists must have an a priori story to tell about how physical nature makes true psychological nature a priori physicalism. Simply put, then, Jackson’s claim is that a priori physicalism is true.

But is a priori physicalism true? That depends on what it means to have an a priori story to tell about the relation between physical nature and psychological nature. In what follows I distinguish two ways in which this claim might be understood. I then argue that on neither way of understanding it does Jackson have an argument for a priori physicalism.

2. I begin with physicalism. Physicalism is the metaphysical thesis that physical nature necessitates psychological nature, or as it is sometimes said, that there can be no psychological difference without a physical difference.

---


2 Jackson is not alone in thinking this; Chalmers (1996) and Lewis (1994) hold similar views.

3 I will say very little about the nature of the physical. In particular, I will say next to nothing about what makes a property a physical property, except to say that by ‘physical property’ I
One way to make this idea more precise is to articulate physicalism as a global supervenience thesis, as follows:

(S1) Physicalism is true iff any two worlds which are physical duplicates are duplicates simpliciter.

However, it is arguable that (S1) is not an adequate definition of physicalism. This is because (S1) is a thesis about all worlds, to the effect that no two worlds can differ without differing physically, whereas physicalism is supposed to be "a merit...that not all worlds share". (Lewis 1983, 210) A restricted supervenience thesis is therefore needed. One such thesis is:

(S) Physicalism is true of a world w iff any physical duplicate of w is a duplicate simpliciter of w.

With respect to the actual world, physicalism will be true just in case any world which is physical duplicate of the actual world is a duplicate

---

simpliciter of the actual world. For the purposes of our discussion, I will assume that (S) is an adequate definition of physicalism.\(^5\)

Physicalism is a significant thesis. But more can be said about it. For it is a consequence of (S) that if physicalism is true of a world \(w\) then the physical nature of \(w\) entails—in a sense to be explained in a moment—the psychological nature of \(w\). To see what this means, suppose \(\phi@\) is the complex physical sentence that states the entire physical nature of the actual world. And suppose \(\psi@\) is the complex psychological sentence that states the entire psychological nature of the actual world. Then if physicalism is true

\[^5\text{It should be granted that this is not an insignificant assumption. This is because it is arguable that (S) will not do as a definition of physicalism either. For it is consistent with (S) that there are worlds which are physical duplicates but which are not duplicates simpliciter. For example, suppose \(w\) and \(w^*\) are physical duplicates, but that \(w^*\) contains floaters, non-physical creatures which are epiphenomenal with respect to the physical nature of \(w^*\) and which enjoy rich psychological lives. (I owe the expression 'floaters' to Judith Jarvis Thomson.) Then although \(w^*\) would be a physical duplicate of \(w\) it is arguable that it would not be a psychological duplicate of \(w\). Or, suppose \(w\) and \(w^*\) are physical duplicates, but that \(w^*\) contains kites, non-physical creatures which interact with—and so 'tug on'—the physical nature of \(w^*\). (I owe the expression 'kite' to Judith Jarvis Thomson.) Then again, although \(w^*\) would be a physical duplicate of \(w\) it is arguable that it would not be psychological duplicate of \(w\) (barring a case of massive overdetermination). The case of kites is complicated by issues about how to individuate physical happenings or events. For example, suppose some physical event \(e\) occurs in both \(w\) and \(w^*\), but that in \(w^*\) \(e\) is caused by the 'tug' of a kite, whereas in \(w\) \(e\) is caused by some physical event. Is \(e\) in \(w\) identical with \(e\) in \(w^*\)? This will depend on whether physical events are individuated in part by their causes. Despite the importance of this issue, however, I cannot go into it in more depth here. At any rate, a patch suggested by Jackson is to replace (S) with (S-) Physicalism is true of a world \(w\) iff any minimal physical duplicate of \(w\) is a duplicate simpliciter of \(w\).\]
of the actual world, it follows that the conditional $\phi_@ \supset \psi_@$ is necessarily true. More generally, if physicalism is true of the actual world it follows that for any true psychological sentence $\psi$ there will be a true physical sentence $\phi$ such that the conditional $\phi \supset \psi$ is necessarily true—that is, if physicalism is true of the actual world every psychological way the actual world is will be made true by some physical way the actual world is. Let us call all necessary truths of the form $\alpha \supset \beta$ fixing conditionals. And let us say that whenever we have a fixing conditional, its antecedent entails—in the necessary truth-preserving sense—its consequent. The claim that

---

6 This raises the issue of contingency. The issue is tricky. For if supervenience is a necessitation relation between sets of properties, how can a supervenience thesis be contingent? But consider: again, let $\phi_@$ be the sentence which expresses the complete physical nature of the actual world, and let $\psi_@$ be the sentence that expresses the complete psychological nature of the actual world. Then if physicalism is true of the actual world any world at which $\phi_@$ is true is a world at which $\psi_@$ is true. This much is uncontroversial. Consequently, when it is said that physicalism is contingent this clearly cannot mean that there could be a world $w$ at which $\phi_@$ is true but $\psi_@$ false, for then $\phi_@ \supset \psi_@$ would not be necessarily true. But then in what sense is physicalism contingent? It is contingent in the following sense. According to (S) a world $w$ is a physicalistic world just in case the physical nature of $w$ necessitates the psychological nature of $w$. Note, however, that the sentence which expresses the physical nature of a given world will differ from world to world, as will the sentence which expresses the psychological nature of a given world. Physicalism is therefore contingent in that it is true of some worlds $w$, but not others, that the sentence that expresses the physical nature of $w$ necessitates the sentence that expresses the psychological nature of $w$. In short, for any world $w$, it is contingent that physicalism is true at $w$; but if physicalism is true at $w$, then the physical nature of $w$ necessitates the psychological nature of $w$.

Moreover, the issue is not restricted to discussions of physicalism. For there are many metaphysical theses that are contingent. Consider, for example, Determinism, understood as the thesis that the state of the world at a time $t$, together with the laws of nature, completely determines the state of the world at some later time $t'$. It is plausible to think that this thesis is contingent, true at some worlds and false at others. So we need to relativize the thesis of Determinism to a world: Determinism will be true of a world $w$ just in case the state of $w$ at $t$ plus the laws of nature true at $w$ completely determines the state of $w$ at some later time $t'$. (I am indebted here to discussions with Robert Stalnaker and Daniel Stoljar.)

7 Talk of fixing conditionals can be found in Byrne (1993) and Jackson (1994, 1997). See also Yablo (1992) for the distinction between a priori—or conceptual—entailment and metaphysical entailment. This notion of entailment corresponds to what is sometimes called strict implication, where a sentence $\alpha \supset \beta$ (proposition) $P$ is said to strictly imply a sentence (or proposition) $Q$ just in case the corresponding material conditional is necessarily true, i.e., just in
physicalism is true therefore entails the claim that there are fixing conditionals with purely physical antecedents and purely psychological consequents. For simplicity, let us call this the Physical Entailment Thesis:

**Physical Entailment Thesis:**

If physicalism is true, then for every true psychological sentence \( \psi \) there is a true physical sentence \( \phi \) such that the conditional \( \phi \supset \psi \) is necessarily true.

So far, so good. Now everybody—including Jackson—agrees that the Physical Entailment Thesis is true; this much is uncontroversial. What is controversial is Jackson’s further claim that physicalists must have “an a priori story to tell about how the physical way the actual world is [entails] the psychological way the actual world is.” (Jackson 1994, 40) What does it mean to have an a priori story to tell about the relation between physical nature and psychological nature? I suggest that this claim of Jackson’s can be understood in two ways. On the one hand, Jackson is sometimes concerned to argue “that issues to do with conceptual analysis broadly conceived are inevitably central to [physicalism].” (Jackson 1994, 24) It is clear, I think, that Jackson means by this that there must be an analysis of psychological vocabulary in non-psychological terms if physicalism is to be true. This suggests the following thesis, which I will call the Conceptual Analysis Thesis.

---

case it is impossible for \( P \) to be true and \( Q \) false. For a formulation of physicalism along these lines, see Kirk 1996.
Conceptual Analysis Thesis:
If physicalism is true then there is an analysis of the psychological in non-psychological terms.

On the other hand, Jackson is sometimes concerned to argue that if physicalism is true then every psychological way the world is is made true a priori by some physical way the world is. More exactly, he is concerned to argue that "physicalism is committed to the a priori deducibility of the psychological from the physical." (Jackson 1997, Lecture 3, 31) Call this the A Priori Entailment Thesis.

A Priori Entailment Thesis:
If physicalism is true then for every true psychological sentence $\psi$ there is a true physical sentence $\phi$ such that the fixing conditional $\phi \supset \psi$ is a priori.$^8,9$

Let me make a few comments about the A Priori Entailment Thesis. The Physical Entailment Thesis says that if physicalism is true there are fixing conditionals with physical antecedents and psychological consequents. The A Priori Entailment Thesis says that if physicalism is true, such fixing conditionals are a priori. To say that a fixing conditional $\alpha \supset \beta$ is a priori is to say that $\beta$ can be deduced from $\alpha$ using logical or conceptual knowledge.

---

$^8$ I assume that the conjunction of two or more physical sentences is itself a physical sentence. Thus, although the antecedent of some fixing conditional $\phi \supset \psi$ might be a complex conjunctive sentence it will still count as a physical sentence for our purposes.

$^9$ The A Priori Entailment Thesis is very close to what Byrne (1998) calls the thesis of Cosmic Hermeneutics. According to Byrne, cosmic hermeneutics is possible "iff, for every true $\psi$ there is some true physical sentence $\phi$ such that $\phi \supset \psi$ is knowable a priori." (2)
alone. In general, any logical truth stated in the form of a conditional will count as an a priori fixing conditional. But we also want to leave it open that there can be a priori fixing conditionals which are not logical truths. An example of such a conditional might be ‘If an object $x$ is red, then $x$ is colored.’ This is not a logical truth, but it is arguably a priori. Consequently, talk about ‘deducing the consequent of a conditional from its antecedent using logical or conceptual acumen alone’ is not restricted to talk about ‘deducibility in some formal system’. In short, the A Priori Entailment Thesis is the conjunction of the claim that if physicalism is true, then for every true mental sentence $\psi$ there is a true physical sentence $\phi$ such that $\phi \overrightarrow{\psi}$ is necessarily true—the Physical Entailment Thesis—with the claim that $\psi$ can be deduced from $\phi$ using logical or conceptual knowledge alone.

In the remainder of this paper I wish to consider Jackson’s arguments for each thesis. I will argue that none of his arguments for the Conceptual Analysis Thesis is compelling. I will also argue that there are no good arguments for the A Priori Entailment Thesis, and good arguments against it.

3. Before turning to Jackson’s arguments for the Conceptual Analysis Thesis, however, two questions need to be addressed. First: what is conceptual analysis according to Jackson? And second: is there any reason to believe that Jackson’s picture of conceptual analysis is tenable? Traditional accounts of conceptual analysis typically hold that an analysans $A$ of an analysandum $C$ is a good analysis if ‘$A$ is $C$’ is a priori, and $A$ states necessary and sufficient conditions for the application of $C$. Jackson thinks of conceptual analysis in a somewhat more liberal manner, however, and attributes to it four main features.
First, responding to an objection of Stephen Stich’s that we cannot presently give an analysis of animal grooming behavior in physical terms, and will probably never be able to do so, Jackson remarks that “[w]hat we require from physicalists who accept the existence of grooming behavior is enough by way of conceptual analysis to make it plausible that the purely physical account of our world makes true the grooming-behavior account of our world; and to do that it is not required to give necessary and sufficient conditions in physical terms for grooming behavior.” (Jackson 1997, Lecture 3, 7-8) While this passage is consistent with a view on which conceptual analysis does provide necessary and sufficient conditions for the application of a concept C, Jackson’s point is that necessary and sufficient conditions are not needed in order to establish that C can be made sense of in purely physical terms.

Second, Jackson argues that conceptual analysis has an empirical element, in that “conceptual analysis in our sense is of a kind with what cognitive psychologists do when they investigate the young child’s concept of faster than, and political scientists do when they investigate different voters’ concept of socialist, and these are, of course, empirical investigations.” (Jackson 1997, Lecture 2, 23)

Third, Jackson argues that his employment of conceptual analysis is modest in nature. What does this mean? I’ll borrow an example from Jackson to illustrate.10 Four-dimensionalism is the view that physical objects have temporal parts. Hence according to four-dimensionalism, a physical object’s having different temperatures, say, is a matter of different temporal parts of the object having different temperatures. But there is an intuition

10 Jackson bases his discussion on that of Geach (1972).
that this is not genuine change, since genuine change requires the same
object having different properties at different times. This is a piece of
conceptual analysis. Here, however, conceptual analysis is being used
modestly to highlight the fact that the folk conception of change differs from
the four-dimensionalist’s conception of change. But this is not to make “any
claim, one way or the other, about what the world is like; [the] claim is
simply that if four-dimensionalism is true, it is right to say that nothing
changes in the folk sense of change.” (Jackson 1997, Lecture 2, 18)

However, consider the following argument:

**Four-Dimensional Argument**

(D1) Different objects having different properties is not change.
   (Conceptual claim.)

(D2) Things change. (Moorean fact.)

(DC) Four-dimensionalism is false. (Claim about the nature of our world.)

This argument is an example of conceptual analysis in its immodest role,
since the fact that the folk conception of change and the four-
dimensionalist’s conception of change are different is being used in an
argument for the conclusion that four-dimensionalism is false as a claim
about the nature of our world. Jackson rejects this immodest role for
conceptual analysis, arguing instead that the role of conceptual analysis “is
that of addressing the question of what to say about matters described in one
set of terms given a story about matters in another set of terms. Conceptual
analysis is not being given a role in determining the fundamental nature of
our world; it is, rather, being given a central role in determining what to say
in less fundamental terms given an account of the world stated in more fundamental terms." (Jackson 1997, Lecture 2, 19-20)\textsuperscript{11}

In Jackson's view, then, conceptual analysis has at least three central features. First, an analysans A of some analysandum C need not state necessary and sufficient conditions for the application of C. Second, conceptual analysis can have a significant empirical element. And third, conceptual analysis is modest in its goals. In addition, conceptual analysis has a fourth feature: it yields a priori results. To support this claim, Jackson employs a specific semantic framework for representing the relationship between meaning and context. This is the so-called two-dimensional framework.\textsuperscript{12} Jackson's central claim is that it follows from the adoption of the two-dimensional framework that there is an a priori component to the meaning of every expression. I will argue that this claim is false and hence, that Jackson's picture of conceptual analysis is mistaken.

\textsuperscript{11} Jackson remarks that the modest role he gives to conceptual analysis has led him to have doubts about the Knowledge Argument (Jackson 1982, 1986). The Knowledge Argument concerns Mary, a brilliant scientist who is locked in a black-and-white room, and who comes to learn—and so know—everything physical there is to know about vision. When she is let out of her room and sees a red rose for the first time, however, it is very plausible to suppose that Mary learns a new fact, for she learns what it's like to see red. And according to the Knowledge Argument, what this shows is that physicalism is false, since it shows that some facts are not physical facts.

What Jackson has in mind, I think, when he says that the modest role given to conceptual analysis has led him to have doubts about the Knowledge Argument is that what the Knowledge Argument shows is that the folk concept has a new experience or learns a new fact is different from the physicalist's concept has a new experience or learns a new fact. Employing conceptual analysis in its modest role, the conclusion to be drawn from this is simply that the two concepts of what it is to learn a new fact differ. Employing conceptual analysis in its immodest role, however, the conclusion to be drawn is that physicalism is false of the actual world. Unfortunately, I cannot discuss this issue in more detail here.

\textsuperscript{12} For expositions of two-dimensionalism see, for example, Stalnaker (1978), Davies and Humberstone (1980), Tichy (1983), and Lewis (1994).
4. The two-dimensional framework takes as its starting point the commonplace observation that there are two ways in which what is said depends on context. First, context determines what the semantic value of an expression is on a given occasion of use. Here context can be said to be functioning in its role as the context of utterance.\(^{13}\) So, for example, given what the world is like, the English sentence ‘water fills the oceans’ is true if, and only if, \(\text{H}_2\text{O}\) fills the oceans. In other words, context determines that the sentence ‘water fills the oceans’ expresses the proposition that \(\text{H}_2\text{O}\) fills the oceans.

But what is said depends on context in another way. For whether what is said is true depends on what the world is like. Here context can be said to be functioning in its role as the context of evaluation. So, for example, ‘water fills the oceans’ is true because, taking the actual world as the context of utterance, ‘water fills the oceans’ expresses the proposition that \(\text{H}_2\text{O}\) fills the oceans, and taking the actual world as the context of evaluation, it is true that \(\text{H}_2\text{O}\) fills the oceans.

For the purposes of this discussion, let us follow Jackson and stipulate that a substance \(S\) is the watery stuff in a world \(w\) if, and only if, \(S\) is a clear, colorless, odorless liquid in \(w\). Let \(@\) be the actual world, in which \(\text{H}_2\text{O}\) is the watery stuff and in which \(\text{H}_2\text{O}\) fills the oceans; let \(w_1\) be a world alike \(@\) in every respect except that in \(w_1\), XYZ is the watery stuff and XYZ fills the oceans; and let \(w_2\) be a world exactly like \(@\) in every respect except that in \(w_2\), \(\text{H}_2\text{O}\) is the watery stuff and Zima fills the oceans.\(^{14}\) Then we can represent

\(^{13}\) This terminology is taken from Kaplan (1989), although I use it in a slightly different way than he does.

\(^{14}\) I have omitted other possible worlds for ease of exposition. Also, Zima is a clear, colorless, alcoholic beverage that is sometimes found in bars around here.
the proposition expressed by ‘water fills the oceans’, with @ understood to be the context of utterance, as follows:

<table>
<thead>
<tr>
<th></th>
<th>@</th>
<th>w₁</th>
<th>w₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>@</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

This represents the fact that the proposition expressed by ‘water fills the oceans’, understood as uttered in @, is true at @ but false at w₁ and w₂. Call this the horizontal proposition expressed in @ by ‘water fills the oceans’.

We can fill out this example further by asking what proposition would ‘water fills the oceans’ have expressed had w₁ or w₂ been the context of utterance. This question is intended to be understood in the same sense in which the answer to the question, What proposition would ‘I am the President’ have expressed if that very sentence had been uttered by me?, is that AB is the President. On this way of understanding it, the answer to the question, What proposition would ‘water fills the oceans’ have expressed had w₁ or w₂ been the context of utterance?, determines the following two-dimensional matrix:¹⁵

---

¹⁵ Again, other possible worlds have been omitted for ease of exposition.
The fact that the horizontal rows following @ and \( w_2 \) are the same indicates that ‘water fills the oceans’ expresses the same proposition when either @ or \( w_2 \) is taken to be the context of utterance. The fact that none of the vertical rows are the same indicates that the relevant context of evaluation is different in each of @, \( w_1 \), and \( w_2 \). In @ it is \( H_2O \) that fills the oceans; in \( w_1 \) it is XYZ; and in \( w_2 \) it is Zima.\(^{16}\)

We can locate another proposition in our matrix, which following Stalnaker (1978) I will call the diagonal proposition. In the case of ‘water fills the oceans’, this is the proposition that is true at a world \( w \) just in case, when \( w \) is taken to be the context of utterance, the substance that is the watery stuff in \( w \) fills the oceans in \( w \). The diagonal proposition is that proposition which is true at @ and \( w_1 \), and false at \( w_2 \). Intuitively, this is the proposition expressed by the sentence ‘the watery stuff fills the oceans’.

These two ways in which what is said depends on context results in two different notions of meaning. Focusing on sub-sentential expressions, Jackson remarks that

\(^{16}\) It should be emphasized that there is a way of understanding the question, What proposition would ‘water fills the oceans’ have expressed had \( w_1 \) or \( w_2 \) been the context of utterance? according to which the answer is, the proposition which is true at @, and false at \( w_1 \) and \( w_2 \). In other words, if we reject Jackson’s assumption that there are two types of meaning associated with every expression, then the horizontal expression expressed by a sentence will not vary from context of utterance to context of utterance. It should be clear that this is not the interpretation at issue here.
[w]e can think of the various possible particulars, situations, events, or whatever to which a term applies in two different ways, depending on whether we are considering what the term applies to under various hypotheses about which world is the actual world, or whether we are considering what the term applies to under various counterfactual hypotheses. In the first case, we are considering, for each world \( w \), what the term applies to in \( w \), given or under the supposition that \( w \) is the actual world, our world. We can call this the \( A \)-extension of term \( T \) in world \( w \)—‘A’ for actual—and call the function assigning to each world the \( A \)-extension of \( T \) in that world, the \( A \)-intension of \( T \). In the second case, we are considering, for each world \( w \), what \( T \) applies to in \( w \) given whatever world is in fact the actual world, and so we are, for all worlds except the actual world, considering the extension of \( T \) in a counterfactual world. We can call this the \( C \)-extension of \( T \) in \( w \)—‘C’ for counterfactual—and call the function assigning to each world the \( C \)-extension of \( T \) in that world, the \( C \)-intension of \( T \). (Jackson 1997, Lecture 2, 24)

In Jackson’s terminology, every expression has associated with it two extensions and two intensions. The \( C \)-extension of a sentence is what I have been calling the horizontal proposition expressed by a sentence. The \( A \)-extension of a sentence, on the other hand, is the proposition the sentence would have expressed had some other world been considered as the context of utterance.

The distinction between \( A \)-intensions and \( A \)-extensions on the one hand, and \( C \)-intensions and \( C \)-extensions on the other, is relevant to our discussion for the simple reason that, according to Jackson, “[i]t is a term’s \( A \)-extension in any world, and so its \( A \)-intension, that is an \textit{a priori} matter. This is because a term’s \( A \)-extension in \( w \) is its extension given that \( w \) is actual, and so does not depend on which world is in fact actual.” (Jackson 1997, Lecture 3, 1) Jackson is not alone in thinking this; David Chalmers concurs. According to Chalmers, “The primary intension of [an expression], unlike the secondary intension, is independent of empirical factors: the
intension specifies how reference depends on the way the external world turns out, so it does not itself depend on the way the external world turns out.” (Chalmers 1996, 57)¹⁷

Here, then, is an argument for the claim that conceptual analysis yields a priori results. Every expression has associated with it a two-dimensional matrix; so every expression has associated with it both an A-intension and an A-extension. Since the A-intension and the A-extension of an expression are a priori, there is an aspect of the meaning of every expression that is a priori.

It seems to me, however, that this claim is misleading at best. That A-intensions and A-extensions are a priori does not follow from the adoption of the two-dimensional framework without certain additional assumptions, as I will now argue.

Jackson is concerned to argue that there is a kind of meaning that is independent of how the world turns out. Whether one agrees with Jackson will depend in part on how one understands the two-dimensional framework. One way to understand the two-dimensional framework is as a way of representing the way in which meaning depends on external factors. On this view, what is basic is the wide or external component of meaning; the diagonal component of meaning is then defined in terms of the external component. A different way of understanding the two-dimensional framework is to take the diagonal component of meaning as basic and to define the wide or external component of meaning in terms of it. It is clear, I think, that Jackson (and Chalmers) think of the two-dimensional

¹⁷ Chalmers’ primary intensions are Jackson’s A-intensions; his secondary intensions are Jackson’s C-intensions.
framework in this second way. (It is no wonder that Chalmers calls the $A$-intensions and $A$-extensions 'primary', and $C$-intensions and $C$-extensions 'secondary'.)

The first way of understanding the two-dimensional framework would have us consider, first, what the $C$-extension of an expression is with the actual world considered as the context of utterance; then consider what the $C$-extension of the expression would be with another world $w$ considered as the context of utterance; and so on. The $A$-extension of the expression would then be determined by reading the $A$-extension off the two-dimensional matrix so defined. According to Jackson, however, this gets things the wrong way around. He would insist that one first uses the $A$-intension associated with an expression to determine that expression's two-dimensional matrix, and then one reads the expression's $C$-extensions off the matrix thus determined. But how exactly is this supposed to work?

Jackson wants to isolate an aspect of meaning that does not depend on which world is actual. But what does this mean? Perhaps what is meant is this: a token of the word 'water', say, will have the same $A$-intension regardless of what world it occurs in. But this can't be right. For consider a world in which 'water' refers to bathtubs; surely there is such a world, and surely that is a world in which 'water' has a different $A$-intension than it actually has. This suggests that there can be no aspect of the meaning of a word that is completely independent of the way the external world turns out, since what $A$-intension a word has depends on the way the external world turns out. Thus, Jackson must mean something else when he says that the $A$-intension of a word does not depend on how the external world turns out.
Now, it might be thought that it is easy enough to see what Jackson has in mind: given our understanding of the meaning of a term $T$, we are in a position to know how the extension of $T$ depends on context, and so are in a position to know the $\mathcal{A}$-extension of $T$. Thus, if I know the meaning of ‘water’, say, I know how the extension of ‘water’ varies depending on what world is taken to be the context of utterance, and so I know the $\mathcal{A}$-extension of ‘water’. Since I can always ask what ‘water’ would have referred to had another world been actual, then it might seem to follow that there is some aspect of the meaning of ‘water’ that does not vary from world to world, and which can plausibly be regarded as a priori. But for unoriginal reasons I think that this will not do.\textsuperscript{18} For how are we to tell, given a possible world $w$ as argument, whether what the $\mathcal{A}$-intension of a term $T$ picks out in $w$ is something other than what its $\mathcal{A}$-intension picks out in the actual world, or whether $T$ is being used with a different $\mathcal{A}$-intension in $w$? For example, is a world in which ‘water’ refers to bathtubs a world in which ‘water’ has a different $\mathcal{A}$-intension than it actually has, or a world in which the $\mathcal{A}$-intension that ‘water’ actually has picks out bathtubs? As we have seen, there is nothing intrinsic to $T$ that makes it the case that it has the $\mathcal{A}$-intension it has. But this suggests there is no way to tell whether, in thinking counterfactually about some term $T$, we have isolated the right $\mathcal{A}$-intension for $T$, and so no way to tell a priori what the $\mathcal{A}$-intension of $T$ is.

It will likely be objected that it is possible to isolate the $\mathcal{A}$-intension of a term $T$ upon suitable reflection. After all, it is our term. But again, there are problems with this claim. We can individuate terms in (at least) two ways. We can regard a term $T$ as being individuated by its syntactic features

\textsuperscript{18} See Block and Stalnaker (1997).
alone, or we can regard $T$ as being individuated in part by certain of its semantic features. Suppose we individuate $T$ by its syntactic features alone. Then the $A$-intension of $T$ is rendered trivial and uninteresting. Suppose again that there is a world $w$ in which ‘water’ picks out bathtubs. Since we are assuming that ‘water’ is individuated by its syntactic features alone, the $A$-intension of our word ‘water’ includes $w$. Indeed, since ‘water’ could in principle pick out anything, there would appear to be no limit on which worlds which are included in the $A$-intension of ‘water’.

Suppose instead that we choose to individuate terms in part by certain of their semantic features. Then it might be thought that sufficient reflection would reveal a term’s $A$-intension. But consider: we want to know what the $A$-intension of our word ‘water’ is. To determine this we take our word ‘water’, with the meaning it actually has, to a world $w$ and ask what it refers to in $w$. But which aspect of ‘water’’s meaning do we take with us? If we take the $C$-intension with us we get the wrong answer, since a world in which $XYZ$ is the dominant clear drinkable liquid will not be included in the $C$-intension of ‘water’. On the other hand, if we take the $A$-intension of ‘water’ with us, then we are presupposing knowledge of the very semantic property that we are attempting to recover. Consequently, there is reason to doubt that even ideal reflection will be sufficient to tell us what the $A$-intension of a term is.  

In short, what the two-dimensional framework establishes is that if there is an a priori aspect of the meaning of an expression, then that aspect can be represented diagonally. What it does not establish is that if a

---

19 It is worth emphasizing that this is not how Jackson chooses to individuate expressions.

20 Block and Stalnaker (1997) make the same point.
component of the meaning of an expression can be represented diagonally, then the component of meaning so represented is a priori. This suggests that Jackson is mistaken in thinking that conceptual analysis yields a priori results.

5. I the previous section I argued that Jackson’s reasons for thinking that conceptual analysis yields a priori results are mistaken. Let me now turn to Jackson’s arguments for the conclusion that conceptual analysis is of central importance to physicalism. Jackson has two such arguments. According to the first, conceptual analysis is required in order to avoid mysteries or ‘acts of faith’. According to the second, conceptual analysis is required in order to define or identify a subject matter. I will disiss each argument in turn.

Let us take as our point of departure the following passage from Jackson:

the physicalist is committed to each and every psychological statement being made true by a purely physical way our world is. The analytical functionalist has a story about how this could be...The story is a piece of conceptual analysis. Analytical functionalism is defended by the a priori methods characteristic of conceptual analysis. For us the important point is that the physicalist must have some story to tell; otherwise how the purely physical makes psychological statements true is rendered an impenetrable mystery. (Jackson 1994, 32)

Now, if Jackson intends this to be an argument for conceptual analysis, it is an unconvincing one. I say this for two reasons. First, the argument provides no reason to think that physicalists in general are committed to conceptual analysis. For even if conceptual analysis provides the only answer to the making-true question, the absence of such an analysis
would not show the metaphysical thesis of physicalism to be false. At best it would show the metaphysical thesis of physicalism to be mysterious. Thus, a commitment to the metaphysical thesis of physicalism does not go over into an argument for the conclusion that physicalists are committed to conceptual analysis.

Second, and perhaps more importantly, the argument provides no reason to think that those physicalists who do wish to avoid mysteries or acts of faith are committed to conceptual analysis. For since it provides no reason for thinking that conceptual analysis is the only answer to the making-true question, it gives us no reason to suppose that only by endorsing conceptual analysis will mysteries be resolved, or acts of faith avoided.

It will perhaps be pointed out that this passage is not intended as an argument for the Conceptual Analysis Thesis but rather as a challenge to those who eschew conceptual analysis but endorse physicalism. On this view, Jackson can be interpreted as saying that conceptual analysis provides one answer to the making-true question, and so should be taken seriously by physicalists. Note, however, that this is only a challenge: it is not an argument that there is no other answer to the making-true question. And even if Jackson thinks that conceptual analysis provides the only answer to the making-true question, it seems clear that the above passage contains no argument for that conclusion.21 In consequence, it remains an open

21 Possibly David Lewis thinks this. As he says,

Arbiters of fashion proclaim that analysis is out of date. Yet without it, I see no possible way to establish that any feature of the world does or does not deserve a name drawn from our traditional psychological vocabulary. (Lewis 1994, 415 [italics mine])

So far as I can tell, however, Lewis, like Jackson, does not provide a specific argument for this conclusion.
question whether physicalists should be overly worried by Jackson’s challenge.

6. Jackson, however, has a second reason for thinking that conceptual analysis is of central importance to physicalism. Says Jackson:

Serious metaphysics requires us to address when matters described in one vocabulary are made true by matters described in another. But how could we possibly address this question in the absence of a consideration of when it is right to describe matters in the terms of the various vocabularies? And to do that is to reflect on which possible cases fall under which descriptions. And that in turn is to do conceptual analysis. Only that way do we define our subject—or rather, only that way do we define our subject as the subject we folk suppose is up for discussion. (Jackson 1997, Lecture 2, 17)

Again, this is perhaps not so much an argument for conceptual analysis, but rather what Jackson calls a ‘rationale’ for it. Still, an argument is suggested by these remarks. I will call this the Defining the Subject Argument.

---

22 A serious metaphysics is a metaphysics that is both discriminative and complete. A metaphysics is discriminative just in case it operates with some limited set of fundamental facts. A metaphysics is complete just in case it attempts to give an account of all (putative) facts in terms of those more limited set of fundamental facts. See Jackson (1994, 1997).
**Defining the Subject Argument**

Premise:
(a) In order to do metaphysics one must address the 'making-true' question, viz., how do matters framed in one vocabulary make true matters framed in another vocabulary?

Premise:
(b) In order to address the 'making-true' question for entities of kind $K$ one must determine what $K$'s are.

Premise:
(c) In order to determine what $K$'s are one must do conceptual analysis.

Therefore:
(d) In order to do metaphysics one must do conceptual analysis.

Unlike the previous argument of Jackson's, the Defining the Subject Argument cannot be dispensed with so easily. For there is something to the idea that in order to say whether a $J$ is a $K$ one must appeal to intuitions about what it is to be a $J$, and that to do this is to do conceptual analysis. Nonetheless, I am not convinced that the Defining the Subject Argument is sound. The problem has to do with (c), and in particular, with the claim that in order to determine or identify a subject matter one must do conceptual analysis.

Now, it might be thought that Jackson's picture of conceptual analysis is sufficiently liberal that virtually any way of identifying a subject matter constitutes an analysis of that subject matter. But this seems to me to be a
mistake. For it is arguable that people can talk about K’s without having an analysis of K-hood. For example, people can surely talk about water, or atoms, or cats, and have only an imperfect understanding of their nature. And this suggests that we can identify a subject matter without doing conceptual analysis.

Still, it might be argued that although the folk needn’t have an analysis of ‘water’ in order to talk about water, there must be such an analysis available, since otherwise it would be impossible to say what ‘water’, for example, refers to. As Jackson says, “any view about how ‘water’ get to pick out what it does will be controversial. But it is incredible that there is no story to tell. It is not a bit of magic that ‘water’ picks out what it does pick out.” (Jackson 1994, 42, n. 23) Jackson is right, of course: there must be some story to tell about how ‘water’ comes to pick out what it does pick out. But does this support his claim that physicalists are committed to thinking that subject-definition requires such analyses in general? Let me sketch some reasons for thinking that it does not.

Suppose one adopts a causal theory of reference. Then there is little reason to believe that interesting analyses of particular words will be forthcoming. 23 For according to most versions of the causal theory of reference also say that they oppose conceptual analysis. But the causal theory of reference is a piece of conceptual analysis in our sense. It is defended by a priori reflection on, and intuitions about, possible cases in the manner distinctive of conceptual analysis. (Jackson 1994, 33)

Even if this is true, however, it is not to the point. What is at issue is whether proponents of the causal theory of reference are committed to thinking that, on the causal theory of reference, the meanings of individual expressions are a priori knowable, or can be given a priori analyses. The claim that the causal theory of reference is piece of conceptual analysis—even if it is true—does not entail this further claim about the meanings of individual words.
reference, the references of words are fixed by the way the world turns out, and not by any conditions that speakers associate a priori with the words themselves. Perhaps on the causal theory of reference it will be a priori that ‘water’ picks out the stuff that plays the right role in causing our ‘water’-thoughts and our water-directed behavior. But even if this it true, it doesn’t distinguish the semantic story for ‘water’ from the semantic story for any other liquid or, indeed, for any other substance.

Second, if defining a subject matter simply involves distinguishing K’s from non-K’s, then it is not implausible to suppose that such subject-defining can be done using reference-fixing descriptions alone. For as Kripke has argued, we can use a description to fix the reference of an expression even if the description isn’t synonymous with the expression whose reference is being fixed. So, for example, I can use the description ‘the President’ to fix the reference of ‘Bill Clinton’ without being committed to thinking that ‘Bill Clinton’ is synonymous with ‘the President’, or to thinking that I have in that manner provided an analysis of ‘Bill Clinton’. More generally, if it is possible to determine when it is right to call something a K by using a reference-fixing description rather than by providing an analysis of K-hood, this would suffice to falsify (c).

Moreover, it seems clear that it is possible to determine when it is right to call something a K by using a reference-fixing description. Take the case of ‘water’, for example. It seems clear that without being in a position to provide necessary and sufficient conditions for the application of ‘water’ I can perfectly well determine when it is right to call something ‘water’, since I can say that something is water just in case it bears the right relation to the
clear, odorless liquid around here. But since the reference-fixing description ‘the clear, odorless liquid around here’ is arguably not an analysis of water, this suggests that (c) is false.

7. The preceding argument made essential appeal to the causal theory of reference. But can Jackson’s picture of conceptual analysis, and the Defining the Subject Argument, really be defeated so easily?

Jackson remarks that his account of conceptual analysis is compatible with the causal theory of reference. This is due to his view that reference-fixing descriptions can be associated with an expression a priori, and that this is all that is needed for the Defining the Subject Argument to go through. Two reasons can be advanced in support of the claim that reference-fixing descriptions might be associated with an expression priori. First, some remarks of Kripke’s suggest that certain natural kind terms have reference-fixing descriptions associated with them a priori. Indeed, Jackson attributes to Kripke the view that “sentences of the form ‘K is...’ where the ellipsis is filled with what reference fixes to K, are a priori[.]” (Jackson Lecture 1997, Lecture 3, 1) Consequently, Jackson could argue that even if we aren’t guaranteed that there is an analysis of ‘water’, say, or ‘pain’, the reference of ‘water’ may yet be fixed by a description that is associated with ‘water’ a priori and that that is sufficient for the Defining the Subject Argument to succeed.

Second, there is Gareth Evans’ example of ‘Julius’.24 ‘Julius’, let us suppose, was introduced into our language as an abbreviation for the definite description ‘the actual inventor of the zip’. Then since the reference

---

24 See Evans (1979) and Evans (1982), 31.
of ‘Julius’ is fixed a priori by the description ‘the actual inventor of the zip’, anyone who understands ‘Julius’ knows a priori that if anyone invented the zip, then Julius invented the zip. So here we have a case where a description $D$ is associated with an expression $E$ a priori, even though $E$ does not mean that same as $D$. Evans calls such names ‘descriptive names’.

As against this, however, two things are worth noting. First, it is implausible to suppose that all words have reference-fixing descriptions associated with them a priori in the manner in which ‘Julius’ does. ‘Julius’ is a philosopher’s example, and while it perhaps succeeds in showing that the idea of descriptive names is coherent, it does not show that such names are commonplace. Indeed, descriptive names—with the possible exception of ‘Jack the Ripper’, ‘The Boston Strangler’, and so on—are exceedingly rare in natural language.

Second, it is an open question whether Kripke’s a priori reference-fixing proposal can be extended to terms other than natural kind terms such as ‘heat’, ‘water’, and possibly, ‘pain’. For example, after discussing the case of heat, Kripke remarks that “[o]ther natural phenomena, such as electricity, are originally identified as the causes of certain concrete experimental effects.” (Kripke 1980, 136-7) But it can hardly be knowable a priori (pace David Lewis, perhaps) that electricity is what gives rise to concrete experimental effect $EE$. So even if it is true that ‘heat is what heat-presents’ is true a priori, this does not support the claim that the references of other natural kind terms are fixed in an a priori manner, let alone the more general thesis that the reference of every expression is fixed in an a priori manner.

To be sure, Kripke does say things which suggest the view that Jackson attributes to him. For example, Kripke remarks that “[i]n the case of
a natural phenomenon perceptible to the senses, the way the reference is picked out is simple: ‘Heat = that which is sensed by sensation S’. Once again, the identity fixes a reference: it therefore is *a priori*” (Kripke 1980, 136). This does suggest a view according to which, whenever we have an identity ‘E = RF’, where RF is the description by which reference to E is fixed, then RF is associated with E a priori. But I think that on reflection this cannot be what Kripke had in mind.

To see why, imagine a case of initial baptism. I am walking along a beach, and see a point of land off in the distance. I say, “Let ‘Andrewsland’ be the name of that point of land with two pine trees and a big rock.” Now, it is arguable that ‘Andrewsland = the point with two pine trees and big rock’ is a priori, at least for me, and at least on that occasion. But it is not plausible to suppose that ‘Andrewsland = the point with two pine trees and a big rock’ is a priori for every competent language user, since different language users will associate different reference-fixing descriptions with what I have dubbed ‘Andrewsland’.

In short, conceptual analysis, while perhaps sufficient for defining or identifying a subject matter, is not necessary. First, it is not implausible to suppose that a subject matter can be identified using reference-fixing descriptions instead of conceptual analyses. And second, even if it is allowed that reference-fixing descriptions are sometimes be associated with terms a priori—witness the example of ‘Julius’, or ‘heat’—there is little reason to suppose that this is true in general. It therefore seems to me that the Defining the Subject Argument does not succeed in establishing the Conceptual Analysis Thesis either.
8. I have argued that Jackson has provided no compelling arguments for the conclusion that physicalists are committed to the Conceptual Analysis Thesis. It remains an open question, however, whether the A Priori Entailment Thesis

A Priori Entailment Thesis:
If physicalism is true then for every true psychological sentence $\psi$ there is a true physical sentence $\phi$ such that the fixing conditional $\phi \supset \psi$ is a priori.

is true, since prima facie at least, it seems that the A Priori Entailment Thesis could be true even if the Conceptual Analysis Thesis is false. It is to this question that I now turn. I begin with a discussion of the problem of a posteriori necessity. I then turn to an argument of Jackson's for the conclusion that physicalists are committed to the A Priori Entailment Thesis and argue that it fails.

It is plausible to suppose that the sentence 'If water is $L$-located, then $\text{H}_2\text{O}$ is $L$-located' expresses a necessary truth. But it seems that we can understand this sentence without knowing that it expresses a necessary truth. So it might seem that understanding a sentence cannot be the same as knowing the conditions under which that sentence is true. What do we understand, then, when we understand the sentence 'If water is $L$-located, then $\text{H}_2\text{O}$ is $L$-located'? According to Jackson, what we understand is the way in which the proposition that the sentence expresses varies depending on context.25 For example, the sentence 'He is the President' can express

25 See also Chalmers (1996) and Lewis (1994).
different propositions in different contexts, depending on who the referent of 'he' is. Similarly, according to Jackson, the sentence 'If water is L-located, then H\textsubscript{2}O is L-located' can express different propositions in different contexts, depending on whether H\textsubscript{2}O or XYZ is the referent of 'water'.\textsuperscript{26} The idea, then, is that what we understand when we understand a sentence like 'If water is L-located, then H\textsubscript{2}O is L-located' is the way in which the content of the sentence varies with context.

With this distinction between meaning and context in hand, Jackson sketches an argument for the conclusion that the psychological nature of the actual world is a priori entailed by the physical nature of the actual world. Consider some true psychological sentence $\psi$. Says Jackson,

understanding and logical acumen are enough to yield how the proposition expressed [by $\psi$] depends on context. But, of course, the context is, according to the materialist, entirely physical. The context concerns various matters about the nature of the actual world, and that nature is capturable in entirely physical terms according to the materialist. Hence, the materialist is committed to there being an \textit{a priori} story to tell about how the physical way things are makes true the psychological way things are. But the story may come in two parts. It may be that one part of the story says which physical way things are, $\Theta_1$, makes some psychological statement true, and the other part of the story, the part that tells the context, says which different physical way things are, $\Theta_2$, make it the case that it is $\Theta_1$ that makes the psychological statement true. What will be a priori accessible is that $\Theta_1$ and $\Theta_2$ together make the psychological statement true. (Jackson 1997, 491)\textsuperscript{27}

\textsuperscript{26} Or: if an XYZ-world had turned out to be actual, the sentence 'If water is L-located, then H\textsubscript{2}O is L-located' would have expressed the necessarily false proposition. Given that the actual world is an H\textsubscript{2}O-world, the proposition that the sentence 'If water is L-located, then H\textsubscript{2}O is L-located' actually expresses—where 'actual' rigidly designates our world—is fixed.

\textsuperscript{27} For a similar argument, see Lewis (1994).
Now, there is a way to understand this argument according to which it is valid, and uncontroversially so. Consider the following:

**Argument A**

(1) If physicalism is true, every sentence is a physical sentence.
Therefore:

(2) If physicalism is true, every sentence is a priori entailed by a physical sentence, namely by itself.

However, while Argument A is clearly valid, it is not to the point. For although no physicalist should deny that, if \( \phi \) is a physical sentence, then \( \phi \) entails \( \phi \) a priori, this is not the conclusion that Jackson is after. Rather, Jackson’s idea is this. Take any true physical sentence \( \phi \) that a posteriori entails some psychological sentence \( \psi \). By hypothesis, understanding \( \phi \) will not be sufficient to enable somebody to move a priori to \( \psi \). But if physicalism is true there will be some additional physical information—information about how the proposition expressed by \( \psi \) depends on context—which, together with \( \phi \), will enable somebody to move a priori to \( \psi \). So if Jackson is right, every psychological way the actual world is made true a priori by some physical way the actual world is.\(^{28}\)

By way of illustration, suppose it is necessarily true, but knowable only a posteriori, that if \( x \) is in a brain state of type \( B \) then \( x \) is in pain.\(^{29}\) Then the following is a valid argument in the modal sense:

---

\(^{28}\) For a slightly different interpretation of Jackson’s argument see Byrne (1998), 31-8.

\(^{29}\) For simplicity I will assume that a brain state of type \( B \) is, in the terminology of Shoemaker (1981), a *total realization* of pain, i.e., is “unconditionally sufficient” for being in pain. This is done to avoid the objection that \( x \)’s being in a brain state of type \( B \) necessitates \( x \)’s being in pain.
Argument B

(3) \( x \) is in a brain state of type \( B \).
Therefore:

(4) \( x \) is in pain.

In other words, it is impossible for (3) to be true but (4) false. By hypothesis, then, (3) entails (4), but only a posteriori. According to Jackson, however, if physicalism is true, there is some additional physical information which, when added to (3), would enable us to move a priori to (4). Let us abbreviate the complicated physical story—whatever it turns out to be—about the nature of pain by simply calling it ‘the pain role.’ Perhaps the details of this physical story will be spelled out in some complicated neurophysiological theory; perhaps they will be filled out with folk platitudes instead. However the story gets filled out, it will in any event be true that a subject \( x \) is in pain just in case \( x \) is in the state that plays the pain role.\(^{30}\) (Talk about ‘filling out’ the pain role in physical terms is intended to be vague. We will consider in a moment how to make it more precise.)

Suppose for the sake of argument that the sentence ‘\( x \) is in pain iff \( x \) is in the state that plays the pain role’ is a priori. And suppose that brain state \( B \) is the state that plays the pain role. Then we could transform the modally valid but a posteriori Argument B into the modally valid yet a priori Argument C:

\(^{30}\) I make the simplifying assumption that there is a unique physical state that plays the pain role.
Argument C

(3) \( x \) is in brain state \( B \).

(3a) Brain state \( B \) is the state that plays the pain role.

Therefore:

(4) \( x \) is in pain.

given the a priori

(3b) \( x \) is in pain iff \( x \) is in the state that plays the pain role.

In short, Jackson's claim is that knowledge of the meaning of '\( x \) is in pain', combined with the context-giving statements (3) and (3a), allows us to move a priori to the conclusion that \( x \) is in pain. Says Jackson,

[t]he crucial point here is the way that the contextual information, the relevant information about how things actually are...enables us to move a priori from the [brain state \( B \)] way things are to the [pain] way they are. But if physicalism is true, all the information needed to yield the propositions being expressed about what the actual world is like...can be given in physical terms, for the actual context is givable in physical terms according to physicalism. Therefore, physicalism is committed to the a priori deducibility of the psychological from the physical. (Jackson 1997, Lecture 3, 31)

This is what we have been calling the A Priori Entailment Thesis.

Is this a compelling argument? To determine whether it is, we first need to say something about Jackson's claim that if physicalism is true 'the actual context is givable in physical terms'. And to do that we need to spend
some time saying what distinguishes physical expressions from psychological expressions.

To a first approximation, a physical expression is an expression that forms part of the vocabulary of the natural sciences such as physics, chemistry, biology, and neurophysiology. A psychological expression, on the other hand, is an expression that forms part of our folk-psychological discourse. So expressions such as 'belief', 'pain', 'desire', and so on, will count as psychological expressions. What remain are expressions which are neither physical nor psychological. Examples of these expressions are 'is typically caused by', 'is the typical cause of', and so on. Let us call these expressions functional or topic-neutral expressions, and let's allow the physicalist to add these expressions to her stock of physical expressions. Then the physical vocabulary will include physical expressions and topic-neutral expressions, but no psychological expressions. These distinctions are obviously rough, but they will suffice to make the points I wish to make.

A number of assumptions have been made in our reconstruction of Jackson's argument for the conclusion that (4) is a priori entailed by (3) and (3a). In particular, it has been assumed that the information expressed by (3), (3a), and (3b) is givable in physical terms. That is, it has been assumed that (3), (3a), and (3b) contain only physical vocabulary and are therefore physical sentences. But prima facie at least, it is doubtful that either (3a) or (3b) is a physical sentence. After all, both contain the expression 'pain', and by hypothesis 'pain' is a psychological expression. Thus, in order for (3a) and (3b) to count as physical sentences, it must be the case that sentences

---

31 See especially J.J.C. Smart (1959) and Lewis (1966).

32 See Block and Stalnaker (1997), Byrne (1998, 40-1), and Yablo (1998) for similar criticisms.
containing the psychological expression ‘pain’ can be recast in purely physical or topic-neutral terms.

Now, there are two natural ways in which one might recast talk of pain in physical or topic-neutral terms. One way is to provide an a priori definition or analysis of ‘pain’. The other way is to provide necessary and sufficient conditions—in physical or topic-neutral terms—for the application of ‘pain’. Let’s look at each strategy in turn. Suppose for the sake of argument that ‘the pain role’ constitutes an a priori definition or analysis of ‘pain’. If so, then (3b) will be a priori, and uncontroversially so, and so will be a priori entailed by the physical way the actual world is. But of course a posteriori physicalists deny that ‘the pain role’ constitutes an a priori definition or analysis of ‘pain’; their claim is precisely that there are no a priori or conceptual analyses of psychological expressions or concepts. I have already argued that the adoption of the two-dimensional framework employed by Jackson does not entail that such an a priori analysis of ‘pain’, or any other expression, is forthcoming. So absent some further argument for the conclusion that ‘the pain role’ constitutes an a priori definition or analysis of ‘pain’, Argument C begs the question against a posteriori physicalism.

Alternatively, suppose that ‘the pain role’ provides us with necessary and sufficient conditions—in physical or topic-neutral terms—for the application of ‘pain’, but that these conditions are knowable only a posteriori. Then the entailment of (4) by (3) and (3a) will not be a priori, and Argument C will not be a priori valid. For if (3b) is only a posteriori entailed by the physical way the actual world is, then although (3a) and (3b) will count as physical sentences, we will not be in a position to move a priori from (3) and (3a) to (4).
In sum, Jackson equivocates on the phrase ‘givable in physical terms’. It is certainly true that physicalists are committed to thinking that all information supervenes on physical information. Hence, if physicalism is true, and if we are given a specification of the complete physical nature of the actual world, then there is a sense in which we are also given the complete psychological nature of the actual world. But this sense of ‘givable in physical terms’ is not the sense of ‘givable in physical terms’ required by Jackson in order for Argument C to count as a priori valid. Jackson needs it to be the case that (3b) is a priori, i.e., that information about psychological nature supervenes on, or is entailed by, physical nature a priori. But this is supposed to be the conclusion of Argument C, not a premise in it.

A slightly different way of making the same point is as follows. Following Terence Horgan (1984), let us distinguish two senses in which information might be physical. Let us say that information is explicitly physical if it is information which is couched in either physical or topic-neutral vocabulary. And let us say that information is ontologically physical if it is information about a physical event, process, or property. Now, if physicalism is true, and if the information reported by (3b) is explicitly physical information, then Argument C will be a priori valid. For (3b) would then be a physical sentence, and so would be a priori entailed by physical nature. However, it is equally clear that (3b) does not report explicitly physical information, since it contains the psychological expression ‘pain’. (Perhaps it will be argued that we can recast talk about ‘pain’ in physical terms. But then the same problems that were encountered above will arise again. Is the recasting of ‘pain’-talk a priori or only a posteriori knowable?)
On the other hand, if the information reported by (3b) is only ontologically physical information, then the inference from (3) and (3a) to (4) is only a posteriori and so Argument C is not a priori valid. For it if (3b) reports only ontologically physical information then a person could be in possession of the information reported by (3) and (3a) and yet not be in a position to move a priori to (4), since she would not be in possession of the relevant physical information reported by (3b). Either way, Jackson’s argument for the A Priori Entailment Thesis fails.

9. I have argued that Jackson’s claim that physicalists must have an a priori story to tell about how the physical way the actual world is necessitates, or makes true, the psychological way the actual world is can be understood in two different ways. On one way of understanding it, Jackson’s claim is equivalent to what I called the Conceptual Analysis Thesis. I argued, however, that Jackson provides no reason for thinking that the Conceptual Analysis Thesis is true. On the other way of understanding it, Jackson’s claim is equivalent to what I called the A Priori Entailment Thesis. I then argued that Jackson’s argument for the A Priori Entailment Thesis is not compelling. The upshot is that a posteriori physicalism is not threatened by Jackson’s arguments.

A few remarks about a priori physicalism in closing. It is insufficiently noted, I think, just how radical and counter-intuitive the thesis of a priori physicalism is. Jackson’s argument for a priori physicalism is driven in part by skepticism about a posteriori necessity. His argument for the A Priori Entailment Thesis takes as its starting point a puzzle about how a person could understand a sentence S and yet not know the conditions under which S is true. Presented with this puzzle, Jackson proposes that
understanding a sentence S does not require knowing the conditions under which S is true, but requires instead knowing how the proposition expressed by S varies with context. And from this he argues that if physicalism is true, the physical way things are, together with linguistic understanding, ought to a priori entail the psychological way things are.

What is interesting about this is the fact that Jackson’s strategy for solving the puzzle presupposes that understanding alone ought to yield knowledge of truth-conditions; for without such a presupposition there would be no puzzle. When extended to sentences which express necessary truths, however, this presupposition is equivalent to the analytic theory of necessity. If it is indeed Jackson’s view that all necessity is linguistic, then it is no surprise to learn that he thinks that there must be an a priori story to tell about necessary truths in general, and about fixing conditionals with physical antecedents and psychological consequents in particular. Jackson began with a puzzle about the relation between meaning, understanding, and truth-conditions. It seems to me, however, that the position he ends up defending is just as puzzling as the puzzle that gives rise to it.33

\[\text{33 I would like to thank Alex Byrne, Matti Eklund, Robert Stalnaker, Daniel Stoljar, Judith Jarvis Thomson, and members of the MIT Graduate Student Discussion Group for comments on previous versions of this material.}\]
1. Conceivability arguments have traditionally played a prominent role in the philosophy of mind. Descartes' argument for the separation of mind and body is perhaps the most famous example of such an argument, but more recent discussions rely heavily on conceivability considerations as well. Despite their prominence, however, it remains a matter of considerable debate whether conceivability arguments are capable of doing the work they are asked to do. Indeed, it seems to me that on reflection such arguments are a good deal less persuasive than many philosophers take them to be. My aim in this paper is therefore to argue that conceivability arguments do not present problems for a certain attractive and widely endorsed version of physicalism in the philosophy of mind.

There are two issues that I want to address in this paper. First, there is an issue about the relation between conceivability and possibility. In brief: do conceivability considerations provide us with evidence about what is, and what is not, possible? And second, there is an issue about whether so-called zombie worlds are coherently conceivable. The paper is structured with these two issues in mind. I begin with a discussion of the version of physicalism that I am interested in defending. I then consider a conceivability argument due to David Chalmers' for the conclusion that this version of physicalism is false. In the next part of the paper I criticize Chalmers' argument against physicalism. I focus in particular on two aspects of his argument. I first argue

\[1\] See, for example, Kripke (1980) and Chalmers (1996).
that Chalmers misrepresents the relation between conceivability and possibility. I then argue that the conceivability of so-called zombie worlds can be accounted for without having to suppose that such worlds are conceivable. The upshot is that the conceivability of zombies does not present a problem for physicalism.

2. But first, some stage-setting. In *The Conscious Mind*, Chalmers argues for property dualism, the view that certain properties of conscious psychological experiences are not identical with any physical properties. Chalmers' main target is a certain view of the relation between physical nature and psychological nature, a view which I will call *a posteriori physicalism*.

A posteriori physicalism is the conjunction of two theses, one metaphysical, the other epistemological. I'm going to take the following thesis as my target formulation of the metaphysical thesis of physicalism:

(S) Physicalism is true of a world w iff any physical duplicate of w is a duplicate simpliciter of w.

A physical duplicate of a world w is a world which is indistinguishable from w in all physical respects. To say that physicalism is true of the actual world is thus to say that any world which is a physical duplicate of the actual world is a duplicate simpliciter of the actual world.²

² For a fuller discussion of how to best formulate the metaphysical thesis of physicalism see Chapter 2. For similar discussions see Lewis (1983), Byrne (1993, 1998), and Jackson (1994, 1997).
The claim that physicalism is true is significant. But more importantly, it also carries with it a commitment to the idea that the physical nature of the actual world entails the psychological nature of the actual world. To see what this means, suppose $\phi_@$ is the complex physical sentence that states the entire physical nature of the actual world. And suppose $\psi_@$ is the complex psychological sentence that states the entire psychological nature of the actual world. Then if physicalism is true of the actual world, it follows that the conditional $\phi_@ \supset \psi_@$ is necessarily true: any world at which $\phi_@$ is true is a world at which $\psi_@$ is true. More generally, if physicalism is true of the actual world then for any true psychological sentence $\psi$ there is a true physical sentence $\phi$ such that the conditional $\phi \supset \psi$ is necessarily true—that is, if physicalism is true of the actual world every psychological way the actual world is will be made true by some physical way the actual world is. For simplicity I will call this the Physical Entailment Thesis:

**Physical Entailment Thesis:**

If physicalism is true, then for every true psychological sentence $\psi$ there is a true physical sentence $\phi$ such that the conditional $\phi \supset \psi$ is necessarily true.

We are now in a position to state the thesis of a posteriori physicalism more precisely. A posteriori physicalism is the conjunction of two claims. First, the claim that physicalism is true of the actual world, and hence—by the Physical Entailment Thesis—that for every true psychological sentence $\psi$ there is a true physical sentence $\phi$ such that the conditional $\phi \supset \psi$ is necessarily true. And second, the claim that this entailment of psychological nature by physical nature is only a posteriori. A posteriori physicalists
therefore "accept that [psychological nature is] not necessitated a priori by physical [nature], but hold that [it is] necessitated a posteriori by physical [nature]." (Chalmers 1998) In short, a posteriori physicalism maintains that although there is a necessary entailment of psychological nature by physical nature, there is no a priori or conceptual connection between the two. This will become important later on.

A posteriori physicalism strikes me as an attractive thesis. First, its metaphysical component makes sense of the idea that physical nature is in some sense basic, that once the physical nature of the actual world is fixed there is no room for further variation. And its epistemological component accounts for the intuitively compelling idea that there is no conceptual connection between physical nature and psychological nature. This aspect of a posteriori physicalism is controversial, however, and some philosophers have argued to the contrary that an a priori or conceptual connection between physical nature and psychological nature is required if physicalism is to have any hope of being true.³ Although none of these arguments strikes me as persuasive, I will not focus on them here.⁴

What I will focus on instead is an argument of Chalmers' for the conclusion that a posteriori physicalism is false. The argument runs as follows:

³ See, for example, White (1986) and Jackson (1994, 1997).

⁴ See Chapters 1 and 2 for arguments to this effect.
Chalmers' Argument

(P1) In the actual world, there are conscious experiences.

(P2) There is a physical duplicate of the actual world in which the facts about consciousness in the actual world do not hold.

(P3) Therefore, facts about consciousness are further facts about the actual world, over and above the physical facts.

(C) So physicalism is false of the actual world.

Central to this argument is the claim that there is a world which is physically identical with, but psychologically different from, the actual world. Given my assumptions concerning physicalism and the Physical Entailment Thesis, I agree that the argument is valid. The questionable premise is therefore (P2). Chalmers' reason for thinking that (P2) is true is that zombie worlds are conceivable, and so are metaphysically possible. I will argue to the contrary, however, that Chalmers' reasons for thinking that zombie worlds are metaphysically possible are flawed. In order to do this, I need to first talk about conceivability, possibility, and the relation between the two notions.

3. It is natural to take the things which are claimed to be conceivable and possible to be propositions. Chalmers, however, focuses on statements instead. For the purposes of this discussion, I will follow him in doing so. A statement, for Chalmers, is a sentence—or something very similar to a sentence—which has truth-conditions or, equivalently, expresses a proposition. Chalmers distinguishes the conceivability of statements from the conceivability of propositions, or worlds. It is important for Chalmers,

\[5\] See Chalmers (1996), 123.
however, that he be able to move from the conceivability of a statement to
the existence of a possible world corresponding to that statement.
('Corresponding to' is deliberately vague; I’ll attempt to make it a bit more
precise below.) To this end, Chalmers suggests the following: a statement is
conceivable (or conceivably true) if it is true in some conceivable, or logically
possible, world.6,7

This implies that a statement \( E \) is conceivable if there is a logically
possible world in which \( E \) is true. But Chalmers also thinks that if a statement
\( E \) is conceivable, then \( E \) is possibly true, and hence, that there is a logically
possible world in which \( E \) is true. For example, he remarks that “whenever it
is conceivable that [a statement \( E \) is true], there will be a logically possible
world in which [\( E \) is true].” (Chalmers 1998) Putting the two halves together: a
statement \( E \) is conceivable if, and only if, there is a logically possible world in
which \( E \) is true.

---

6 What Chalmers in fact says is that a statement is conceivable if it is true in all conceivable
worlds. (Chalmers 1996, 66) I think it is plain, however, that Chalmers simply misspoke here,
as he elsewhere says things that indicate that he thinks that it is truth in some logically
possible world which is the mark of conceivability. (Witness, for example, his remark that
“[o]n this view of conceivability, the conceivability of a statement involves two things, first
the conceivability of a relevant world, and second, the truth of the statement in that world.”
(1996), 67 [italics mine].)

Also, although Chalmers talks of conceivable, or logically possible, worlds—and uses
the two more or less interchangeably—this should not be taken to mean that he thinks that
there are two distinct sets of worlds, the conceivable or logically possible ones, and the
metaphysically possible ones, and that the latter are a subset of the former. The notion of a
logically possible world functions as a primitive for Chalmers. As he says, “[a]s for the notion
of a logically possible world, this is something of a primitive...we can intuitively think of a
logically possible world as a world that God might have created.” (Chalmers 1996, 66)
However, since any world that God might have created is presumably a possible world, it is
hard to avoid the conclusion that by ‘logically possible world’ Chalmers simply means
‘possible world’.

7 For ease of exposition I will sometimes say that a statement \( E \) is conceivable (or that \( E \) is
possible). This is to be understood as short for: \( E \) is conceivably true (or \( E \) is possibly true).
The interesting part of this biconditional for our purposes is the claim that a statement $E$ is conceivable only if there is a possible world in which $E$ is true. Call this (Con):

(Con) If a statement $E$ is conceivable then $E$ is possibly true.

According to (Con), in other words, if a statement $E$ is conceivable, then there is a possible world in which $E$ is true. I said above that it is important for Chalmers that he be able to move from the conceivability of a statement to the existence of a possible world ‘corresponding to’ that statement. (Con) allows him to do precisely that.

Let us return to Chalmers’ argument against a posteriori physicalism:

**Chalmers’ Argument**

(P1) In the actual world, there are conscious experiences.

(P2) There is a physical duplicate of the actual world in which the facts about consciousness in the actual world do not hold.

(P3) Therefore, facts about consciousness are further facts about the actual world, over and above the physical facts.

(C) So physicalism is false of the actual world.

Again, what is at issue is (P2). Chalmers’ reason for thinking that (P2) is true is that zombies—“someone or something physically identical to me (or to any other conscious being), but lacking conscious experiences altogether” (Chalmers 1996, 95)—are conceivable, and hence, that zombie worlds—worlds inhabited by zombies—are possible. Chalmers’ reason for finding zombies and zombie worlds conceivable is found in the following passage:
the [conceivability] of zombies seems...obvious to me. A zombie is just something physically identical to me, but which has no conscious experience—all is dark inside. While this is probably empirically impossible, it certainly seems that a coherent situation is described; I can discern no contradiction in the description. In some ways an assertion of this logical possibility comes down to a brute intuition but...[a]lmost everybody, it seems to me, is capable of conceiving of this possibility. (Chalmers 1996, 96)

In this passage, Chalmers suggests a test for determining whether a situation, or world, is conceivable. The test seems to be this: if no contradiction can be discerned in the description of a situation, then the situation so described is conceivable. So in particular, zombie worlds will be conceivable if there is no contradiction in the description of a situation, or world, which is physically indistinguishable from the actual world, but in which there are no conscious psychological states. Chalmers offers the following argument for (P2):

Zombie Argument

(Z1) It is conceivable that there is a world which is physically indistinguishable from the actual world but in which there is no consciousness, i.e., it is conceivable that there are zombie worlds.

(Z2) If it is conceivable that there are zombie worlds, then it is possible that there are zombie worlds.

(ZC) It is possible that there are zombie worlds.

It should be granted that this argument has some intuitive force, especially against proponents of a posteriori physicalism. For since according to a posteriori physicalism there is no a priori or conceptual connection
between physical nature and psychological nature, the claim that zombie worlds are conceivable seems initially plausible, as does the claim that if it is conceivable that there are such worlds, then such worlds are possible. Moreover, it is also clear that the conclusion of the zombie argument presents problems for physicalism. For if zombie worlds are possible then physicalism is false of the actual world, as the existence of such worlds will entail the falsity of the claim that for every true psychological sentence $\psi$ there is a true physical sentence $\phi$ such that the conditional $\phi \supset \psi$ is necessarily true. Nonetheless, despite the initial plausibility of Chalmers' zombie argument, I think that there are problems with it. I will argue, first, that Chalmers equivocates on the crucial notion of conceivability and that as a result, his argument is invalid. I will then argue, using some of Chalmers' own terminology, that zombies are not conceivable in the relevant sense, and hence, that (Z1) is false.

4. I'll begin with a discussion of conceivability, and in particular with (Con)—the claim that if a statement $E$ is conceivable, then $E$ is possible. This sort of view about the relation between conceivability and possibility is not without critics. Joseph Levine, for example, argues that "what is [conceivable] is an epistemological matter... It takes another argument to get from [an] epistemological possibility...to [a] metaphysical possibility." (Levine 1993, 544) Or again: "suppose we reject the Cartesian model of epistemic access to metaphysical reality altogether. One's ideas can be as clear and distinct as you like, and nevertheless not correspond to what is in fact possible. The world is structured in a certain way, and there is no guarantee that our ideas will correspond appropriately." (Levine 1993, 544)
One way to interpret Levine’s objection is that there is no route from conceivability to possibility, that the conceivability of a statement $E$ provides no evidence for $E$’s possible truth. On reflection, however, I think it is arguable that this is not a tenable view. I say this because it seems clear that conceivability considerations do sometimes provide us with evidence about what is and what is not possible. For example, it seems to me that it is conceivable that I could exist in the absence of Building 20, and it seems odd to deny that this provides me with evidence for thinking that it is possible that I could exist in the absence of Building 20. To deny that conceivability considerations ever provide us with evidence about what is possible seems implausible. What we want to know is why appeals to conceivability considerations are legitimate in some cases but not in others. To claim that they are never legitimate is to ignore this important question altogether.

Of course, it might be thought that we can provide arguments for the view that the evidence provided by conceivability considerations is specious. Consider, for example, Descartes’ argument for the separation of mind and body. Descartes was concerned to argue that persons are not physical objects. Descartes argued that “from the fact that I have gained knowledge of my existence without noticing anything about my nature or essence except that I am a thinking thing, my essence consists solely in the fact that I am a thinking thing.” (Descartes 1993, 132) And since according to Descartes his body is not something whose essence consists solely in the fact that it is a thinking thing, he concluded that he is not identical with his body. The

---

8 It's not clear to me that this is Levine's view, but it is nonetheless suggested by his remarks. Even if Levine doesn't hold this view, however, I think that it is a view to which many are sympathetic.
general principle at work in Descartes' argument seems to be this: if I do not notice that a property $P$ is part of my essence, then $P$ is not part of my essence.

**Descartes' Principle:**

For all individuals $x$, and all properties $P$, if $x$ does not notice that $P$ is part of $x$'s essence, then $P$ is not part of $x$'s essence.

A number of commentators have taken issue with Descartes' principle. For example, Arnauld asked how it is supposed to follow from the fact that I don't *notice* that a property $P$ is part of my essence that $P$ is *not* part of my essence. Or—in the particular case we are concerned with—how is it supposed to follow from the fact that I don't notice that my essence includes being physically embodied that being physically embodied is not part of my essence? Rejecting Descartes' principle would provide us with an attractive way to rebut his argument for the conclusion that persons are not physical objects.

As Yablo points out, however, rejecting Descartes' principle for Arnauldian reasons results in a highly implausible view about the evidential weight of conceivability intuitions. For example, it seems possible to me that I could have been born on a day other than the day I was in fact born on. If the line of argument suggested by Arnauld is any good, however, this reasoning is suspect. For how do I know, from the fact that I do not notice that the property of being born in Toronto is part of my essence, that that property is *not* part of my essence? (Compare: how do I know, from the fact that I do not not

---

9 See Yablo (1990), 159 for a fuller discussion of Arnauld's objection, and Descartes' responses to it.
notice that there is a mouse in the corner of my room, that there is not a mouse in the corner of my room?) Perhaps, despite my not noticing it, my essence is such that it includes within it the property of being born in Toronto. Without something like Descartes’ Principle in the background, its seeming to me that I could have been born in a different place would seem to give me no reason to believe that I could have been born in a different place, and that seems highly implausible. The problem is perfectly general: rejecting Descartes’ principle makes all modal intuitions suspect, and we have good reason to think that this cannot be right. Consequently, it seems to me to be a mistake to object to (Con) on the grounds that conceivability considerations never provide us with evidence about what is and what is not possible.  

A more subtle objection to the sort of account of the relation between conceivability and possibility proposed by Chalmers is due to James van Cleve (1983). Says van Cleve,

[w]hatever we mean by ‘conceivable’, we shall have to contend with the following objection. In 1742 Goldbach proposed to Euler that every even number greater than two is the sum of two primes—a conjecture that to this day has eluded both counterexample and proof. Now isn’t it conceivable that Goldbach’s conjecture is true? Most people to whom I have put his question say yes. But now consider the negation of Goldbach’s conjecture: isn’t it conceivable as well? Most people to whom I have put the question again say yes. Thus many people profess to find Goldbach’s conjecture and its negation equally conceivable. But one of them is impossible! Since either Goldbach’s conjecture or its negation must be true and since every proposition of mathematics is necessarily true if true at all, one proposition in this pair must be a necessary truth and the other an impossibility. It thus appears that at least one impossible proposition is conceivable. (van Cleve 1983, 36)

\[10\] It might be thought that a commitment to Empiricism provides another reason to suppose that modal intuitions are never legitimate, since according to Empiricism intuitions, not being derived from experience, lack evidential weight. However, I think that this observation counts more against Empiricism than it does against the appeal to modal intuitions in our reasoning. For arguments to this effect see Bealer (1992).
Van Cleve suggests that we distinguish two senses of conceivability—a positive and a negative sense—as follows:¹¹

**Strong Conceivability**

A statement $E$ is *strongly conceivable* for a subject $S$ iff $S$ is able to see that $E$ is possibly true.

**Weak Conceivability**

A statement $E$ is *weakly conceivable* for a subject $S$ iff $S$ does not see that $E$ is necessarily false.

Let us set to one side the question what it is to see that a statement is possible; I will return to it shortly.

At any rate, with the distinction between Strong and Weak Conceivability in hand, van Cleve argues that Goldbach’s conjecture is only weakly conceivable. On this view, when people claim to find Goldbach’s conjecture (or its negation) conceivable, what they mean is that they do not see that Goldbach’s conjecture is impossible (nor do they see that the negation of Goldbach’s conjecture is impossible). However, since Goldbach’s conjecture is necessarily true if it is true at all, we know that either Goldbach’s conjecture or its negation is logically impossible. Thus, since at least one of the statements claimed to be weakly conceivable is necessarily false, we can

---

¹¹ Van Cleve's original definitions quantify over propositions rather than statements, but I have reformulated them to accord with Chalmers’ usage.
conclude that it does not follow from a statement’s being weakly conceivable that it is possibly true.

What about Strong Conceivability?\textsuperscript{12} If $S$ is able to see that $E$ is possibly true, does it follow that $E$ is possibly true? Here things get more complicated. It might be thought that the answer to this question is yes: if $S$ sees that $E$ is possibly true, then it follows that $E$ is possibly true. For if ‘sees’ is understood factively, a subject cannot see that a statement $E$ is possibly true and yet it be the case that $E$ is not possibly true. With ‘ees’ understood factively, the conceivability of a statement $E$ has $E$’s possibility built into it, since on this interpretation the claim that $S$ sees that $E$ is possibly true entails that $S$ knows that $E$ is possibly true and hence, that $E$ is possibly true. If this is how ‘sees’ is meant to be understood, however, it becomes extremely puzzling why anybody ever worried about whether the conceivability of a statement $E$ entails the possible truth of $E$. For since on this interpretation conceivable has possibility built into it, there is no question but that conceivability entails possibility.

What this suggests is that the factive interpretation of Strong Conceivability is not the sense of ‘is conceivable’ that is at issue. Luckily, however, there is another sense of ‘sees’ according to which ‘$S$ sees that a statement $E$ is possible’ means merely that ‘it seems or appears to $S$ that $E$ is possible’, from which it does not follow that $E$ is possible. What does it mean to say that it seems or appears that a statement $E$ is possible? Here I think we can distinguish two ways in which a statement might seem to be possible.

Suppose I am wondering whether I could exist in the absence of Building 20. It might \textsuperscript{12} that this possibility seems extremely plausible to me,

\textsuperscript{12} I am indebted to Robert Stalnaker here.
that it simply strikes me as a genuine possibility. Here ‘sees’ is to be understood in the sense of ‘seems overwhelmingly likely’. I do not propose to offer anything by way of an analysis of this notion; instead, I propose to simply stick with the intuitively plausible idea that some statements simply strike us as being possible.

On the other hand, suppose I am wondering whether I could exist in the absence of Building 20. It might be that in order to determine whether this is conceivable I first vividly imagine Building 20 being destroyed and then check to see whether I still exist after its destruction. Here ‘sees’ is to be understood in the sense of ‘can vividly or coherently imagine’. Again, I do not propose to analyze talk about vivid or coherent imagination.

What is interesting about this is that, combining the positive and negative senses of ‘is conceivable’ suggested by van Cleve with the factive and non-factive senses of ‘sees’, we are left with four different senses of ‘is conceivable’:

**Factive Strong Conceivability**

A statement $E$ is *factively strongly conceivable* for a subject $S$ iff $S$ is able to see that $E$ is possibly true.

**Apparent Strong Conceivability**

A statement $E$ is *apparently strongly conceivable* for a subject $S$ iff it appears to $S$ that $E$ is possibly true.

**Vivid Strong Conceivability**

A statement $E$ is *vividly strongly conceivable* for a subject $S$ iff $S$ is able to vividly imagine that $E$ is possibly true.
Weak Conceivability

A statement $E$ is weakly conceivable for a subject $S$ iff $S$ does not see that $E$ is necessarily false.

Now, it is uncontroversial that Weak Conceivability is a non-starter as a definition of 'is conceivable', since it is consistent with Weak Conceivability that necessary falsehoods are conceivable. So we are left with a choice between Factive Strong Conceivability, Apparent Strong Conceivability, and Vivid Strong Conceivability as definitions of 'is conceivable'. Moreover, it is arguable that Factive Strong Conceivability cannot be an adequate account of what it means to find a statement conceivable, for the simple reason that it has possibility built into it. This leaves us with Apparent Strong Conceivability and Vivid Strong Conceivability as candidate definitions of 'is conceivable'.

What is important for our purposes is that neither the Apparent Strong Conceivability nor Vivid Strong Conceivability of a statement $E$ entails $E$'s possibility. At best, the Apparent Strong Conceivability or the Vivid Strong Conceivability of a statement $E$ provides us with prima facie evidence for the possible truth of $E$.

Let us return to Chalmers' zombie argument. With our two different interpretations of the relation between conceivability and possibility in hand, we can offer the following two interpretations of Chalmers' zombie argument:

---

13 Indeed, Chalmers rejects Weak Conceivability for precisely this reason. See Chalmers (1996), 67.
Zombie Argument (Apparent)

(Z1-A) It is apparently strongly conceivable that there is a world which is physically indistinguishable from the actual world but in which there is no consciousness, i.e., it is apparently strongly conceivable that there are zombie worlds.

(Z2-A) If it is apparently strongly conceivable that there are zombie worlds, then it is prima facie possible that there are zombie worlds.

(ZC-A) It is prima facie possible that there are zombie worlds.

and

Zombie Argument (Vivid)

(Z1-V) It is vividly strongly conceivable that there is a world which is physically indistinguishable from the actual world but in which there is no consciousness, i.e., it is vividly strongly conceivable that there are zombie worlds.

(Z2-V) If it is vividly strongly conceivable that there are zombie worlds, then it is prima facie possible that there are zombie worlds.

(ZC-V) It is prima facie possible that there are zombie worlds.

Both Zombie Argument (Apparent) and Zombie Argument (Vivid) seem to me to be intuitively very plausible. In particular, since it is not implausible to suppose that it is both apparently and vividly strongly conceivable that there are zombie worlds, both (Z1-A) and (Z1-V) seem reasonable enough. The problem, however, is that neither (ZC-A) and (ZC-V)
is the conclusion that Chalmers is after, namely that it is possible that there are zombie worlds. For even if we grant that ‘there are zombie worlds’ is apparently strongly conceivable, say, this only provides us with prima facie evidence for the possibility that there are zombie worlds; something else is needed to bridge the gap between this claim and the claim that zombie worlds are possible.

I earlier suggested that Factive Strong Conceivability is not a plausible interpretation of ‘is conceivable’ on the grounds that it trivializes the debate between those who think that conceivability considerations provide us with evidence about what is and what is not possible, and those who do not. Still, suppose we interpret Chalmers’ zombie argument using Factive Strong Conceivability instead; does the resulting argument fare any better? The argument we get runs as follows:

**Zombie Argument (Factive)**

\[(Z1-F)\] It is factively strongly conceivable that there is a world which is physically indistinguishable from the actual world but in which there is no consciousness, i.e., it is factively strongly conceivable that there are zombie worlds.

\[(Z2-F)\] If it is factively strongly conceivable that there are zombie worlds, then it is possible that there are zombie worlds.

\[(ZC-F)\] It is possible that there are zombie worlds.

Zombie Argument (Factive) is clearly valid: \((ZC-F)\) follows from \((Z1-F)\) and \((Z2-F)\) and, moreover, is the conclusion that Chalmers is after. Nonetheless, I think it is clear that Zombie Argument (Factive) will not do, and for a simple reason. The problem is that although \((Z2-F)\) bridges the gap
between conceivability and possibility—owing, again, to the fact that 'sees' is being interpreted factively—(Z1-F) is false, or so it seems to me: it is simply not the case that is it factively strongly conceivable that there are zombie worlds. I realize that this amounts to nothing more than an assertion at the moment, and I will have much more to say about why I think that (Z1-F) is false below.

In short, the problem with Chalmers' zombie argument is that it equivocates on the phrase 'is conceivable'. The reading of 'is conceivable' according to which Chalmers' desired conclusion follows from the second premise of the argument is not the reading of 'is conceivable' according to which the first premise of the argument is plausible. I conclude that Chalmers' zombie argument is invalid, and so does not support (P2).

5. Thus far I have been concerned to argue that there is a gap between the conceivable truth of the statement 'there are zombie worlds' and its possible truth. To this it might be objected that the claim that zombie worlds are conceivable but not metaphysically possible gives rise to problems for which there are no satisfactory answers, and so should be rejected for that reason. In particular, won't this view lead to skepticism about modal intuitions generally? Let me address this issue briefly before continuing.

I have been arguing that the conceivable truth of a statement $E$ does not imply the claim that $E$ is possibly true. But then how are we to tell, for any conceivably true statement, whether there is a possible world corresponding to it? If we cannot be assured that there is a world corresponding to a statement which is conceivably true, then aren't intuitions about modality and conceivability always going to be suspect? In order to determine whether this criticism is a good one we need an account of modal error. The best
account of modal error that I know of is due to Stephen Yablo. Says Yablo, concerning the evidence provided by the conceivable truth of a statement E,

this prima facie evidence [for the possible truth of E] is defeated if there is not improbably a [statement P] such that (a) [P] is true, (b) if [E] is true, then [P] is impossible, and (c) [E] is conceivable because one was unaware of (a) and/or (b). (Yablo 1992, 254, ft. 24)

In short: the prima facie evidence for the possible truth of 'there are zombie worlds' provided by the conceivability of 'there are zombie worlds' is defeated if there is not improbably a statement P such that (a) P is true, (b) if P is true, then it is not possible that 'there are zombies-worlds' is true, and (c) 'there are zombie worlds' is conceivable because one is unaware of (a) and/or (b). Is it not improbable that there is such a P? Unsurprisingly, I think it is.

Suppose P = 'physicalism is true of the actual world'. Then, I think, the prima facie evidence provided by the conceivability of 'there are zombie worlds' is plausibly defeated. For it is not improbable, first, that physicalism is true of the actual world and second, that the reason zombie worlds seem possible is that we are unaware that physicalism is true of the actual world.

This will no doubt prompt the objection that a similar argument can with equal effectiveness be run to show that the conceivability of zombie worlds defeats the prima facie evidence provided by the conceivable truth of physicalism. For suppose P = 'there are zombie worlds'. Wouldn't its truth defeat the prima facie evidence provided by the conceivability of 'physicalism is true of the actual world'? I do not think that it would. For this objection ignores an important point, namely that there are reasons having nothing to do with its conceivability for supposing that physicalism is true of the actual world. For example, there are many a posteriori reasons, stemming largely
from the successes of the natural sciences in explaining most natural phenomena, for thinking that physicalism is true. And since what is actually true is possibly true, these reasons constitute an argument for thinking that physicalism is possibly true.

In short, I am not concerned to argue that dualism could not possibly be true; I am concerned to argue that conceivability considerations of the sort appealed to by Chalmers do not show a posteriori physicalism to be false (although in the end it does seem to me that the possibility that physicalism is true is more plausible than is the possibility that there are zombie worlds).

6. Where are we? Section 4 established that Chalmers' zombie argument is invalid: on the most plausible interpretation of 'is conceivable', the possibility of zombie worlds does not follow from the conceivability of such worlds. I'm now going to switch gears and argue, using some of Chalmers' own terminology, that zombie worlds are not in fact conceivable. This will suffice to show that Chalmers' zombie argument, in addition to being invalid, is also unsound.

Chalmers' view is that "we need a logically possible world for every ideally conceivable [statement]." (Chalmers 1998) To understand why he thinks this, we need to spend a bit of time discussing the two-dimensional semantic framework employed by Chalmers.14 This is because given the two-dimensional framework there are two senses in which a possible world might correspond to a given statement.

---

14 I discuss the two-dimensional framework in more detail in Chapter 2. For further discussion see also Stalnaker (1978), Davies and Humberstone (1980), Tichy (1983), Lewis (1994), Block and Stalnaker (1997), Jackson (1997), and Byrne (1998).
The two-dimensional framework takes as its starting point the idea that there are two ways to think about the meaning of linguistic expressions. According to two-dimensionalism, associated with every expression are two intensions, or meanings, a primary intension and a secondary intension. An intension is a function from possible worlds to referents. The primary intension of an expression $E$ says what $E$ would refer to in a possible world $w$ had $w$ turned out to be actual.\(^{15}\) The secondary intension of an expression $E$, on the other hand, says what $E$ refers to in any possible world given what $E$ actually refers to. Consequently, when Chalmers says that there must be a possible world corresponding to every conceivable statement, he might mean either of two things. On the one hand, he might mean that there must be a possible world corresponding to a statement conceived of, or evaluated, according to its primary intension; on the other hand, he might mean that there must be a possible world corresponding to a statement conceived of, or evaluated, according to its secondary intension.

For example, using ‘water’ and ‘H\(_2\)O’ as our examples, we can say that ‘water is not H\(_2\)O’ is conceivable according to its primary intension and so possible according to its primary intension, although it is neither conceivable nor possible according to its secondary intension. This is because it is plausible to suppose that ‘water’ and ‘H\(_2\)O’ have the same secondary intension, but different primary intensions. So although there are no worlds in the secondary intension of ‘water’ which are not in the secondary intension of ‘H\(_2\)O’, there are worlds which are in the primary intension of ‘water’ but are not in the primary intension of ‘H\(_2\)O’.

\(^{15}\) Following the terminology of Davies and Humberstone (1980).
I said at the outset that I was interested in two issues. One concerned the question whether conceivability considerations provide us with evidence about what is, and what is not, possible. The other concerned the question whether zombie worlds are coherently conceivable. Of course, it is one thing to argue that, in general, conceivability considerations provide us with defeasible evidence for the possible truth of a given statement; it is another thing to argue, with respect to the particular case of ‘there are zombie worlds’, that such worlds are not conceivable. It is to this issue that I now turn.

In discussing the objection that ‘there are zombie worlds’ could be conceivable despite the fact that no zombie worlds correspond to it, Chalmers remarks that

the only route available to an opponent here is to claim that in describing the zombie world as a zombie world, we are misapplying the concepts, and that in fact there is a conceptual contradiction lurking in the description. Perhaps if we thought about it clearly enough we would realize that by imagining a physically identical world we are thereby automatically imagining a world in which there is conscious experience. But then the burden on the opponent to give us some idea of where the contradiction might lie in the apparently quite coherent description. If no internal incoherence can be revealed, then there is a very strong case that the zombie world is logically possible. (Chalmers 1996, 99)

Is it possible to offer an argument for the conclusion that in describing a world as a ‘zombie world’ we are misdescribing it? Elaborating on an argument originally due to Richard Boyd (1980), I will argue that it is. I want to begin, however, by discussing a slightly different example.

In *Naming and Necessity* Saul Kripke considers a number of consequences of the thesis that proper names and natural kind terms are rigid
designators. One consequence is that certain identities previously thought to be only contingently true turn out to be necessarily true. One such identity is (1):

\[(1) \quad \text{heat} = \text{molecular K.E.}\]

According to Kripke, because the terms flanking the identity sign in (1) are rigid designators, (1) is, if true, necessarily true. Nonetheless, there is an intuition that (1) is only contingently true.

As is well known, Kripke has an explanation of why (1) appears contingent. According to Kripke, before we knew what heat was we knew which sensations it gave rise to. In particular, heat is something we identified by virtue of the fact that it gives rise to a certain sensation, which we call 'the sensation of heat'. This is not to say that 'heat' means 'whatever gives rise to the sensation of heat'. Rather, the sensation of heat is the property by which we initially identified the phenomenon of heat. Thus, when we discovered that heat = molecular K.E., we discovered an identification which gives us an essential property of this phenomenon. We...discovered a phenomenon which in all possible worlds will be molecular motion—which could not have failed to be molecular motion, because that's what the phenomenon [of heat] is. On the other hand, the property by which we identify it originally, that of producing such and such a sensation in us, is not a necessary property [of heat] but a contingent one. (Kripke 1980, 133)

---

16 An expression \(a\) is said to rigidly designate an object \(x\) just in case \(a\) refers to \(x\) in every world in which \(x\) exists, and does not refer to something other than \(x\) in worlds in which \(x\) does not exist. See Kripke (1980), 48-9.
According to Kripke, it is this contingent property of heat that accounts for our intuition that heat might not have been molecular K.E. When (1) is imagined to be contingent—and so false—what is being imagined is a world in which something other than molecular K.E. gives rise to sensations of heat. But this is not a counter-example to (1), since it is not a world in which heat ≠ molecular K.E. Rather, it is a world in which something else has the property—contingently possessed by molecular K.E. in this world—of giving rise to sensations of heat. But since ‘heat’ does not mean ‘that which gives rise to sensations of heat’ such a world is compatible with the truth of (1).

What Kripke offers is a strategy for redescribing the conceivable falsity of certain necessary truths in a way that does not reflect badly on the modal status of those truths. To recapitulate, whenever we have a necessary a posteriori truth, there is always going to be an intuition that the truth is only contingent—that is, because the truth is only a posteriori it will seem conceivably false. So if we want to insist that a certain statement is both a posteriori and necessarily true, we need to redescribe it in a way that accounts for the appearance of contingency but does not reflect badly on its modal status. In the case of (1), we can do this by finding a contingent property of heat via which reference to it was initially fixed, and arguing that what we are imagining when we are imagining that (1) is false is that something other than molecular K.E. has that contingent property.

What does Kripke’s redescription strategy have to do with the two-dimensional framework employed by Chalmers? Simply put, Kripke’s contingent reference-fixing properties are Chalmers’ primary intensions. In general, we can recover the primary intension of an expression containing

\footnote{For a nice discussion and comparison of Chalmers and Kripke on this issue, see Yablo (1998).}
rigid designators by replacing one or more of the rigid designators with the descriptions by which reference was initially fixed. For example, the primary intension of 'heat' is a function which picks out the thing around here that gives rise to sensations of heat. This is because that description represents the way in which we think about the reference of 'heat'. Hence, when evaluated according to the primary intension of 'heat', 'heat is molecular K.E.' becomes 'the thing around here that gives rise to sensations of heat is molecular K.E.'.

The differences between the two-dimensional view advocated by Chalmers and Kripke's reference-fixing proposal therefore seem to be a matter of terminology.¹⁸

Now, what goes for necessary a posteriori identities also goes, mutatis mutandis, for necessary a posteriori conditionals of the sort to which I have argued physicalism is committed. For simplicity, let's focus on a simple conditional linking physical nature and psychological nature, for example,

(2) If $x$ is in a brain state of type $C$ then $x$ is conscious.¹⁹

Following Shoemaker (1981), I will assume that being a brain state of type $C$ is a total realization of being conscious and hence, that brain states of type $C$ are unconditionally sufficient for being conscious. In other words, I will assume that there is no world in which a person is in a brain state of type $C$ but is not conscious owing to the fact that there is some other property that that person lacks, the instantiation of which is required for the instantiation of the

---

¹⁸ Witness Chalmers: a "difference in primary intensions will correspond to a difference in reference-fixing properties." (Chalmers 1996, 373, ft. 15)

¹⁹ This is obviously very much simplified; however, it will suffice to make the point I want to make.
property of being conscious. Hence, on the assumption that physicalism is true of the actual world it follows that (2) is, if true, necessarily true: any world in which the antecedent of (2) is true is a world in which the consequent of (2) is true. However, on the assumption that a posteriori physicalism is true it will seem possible that there could be worlds in which there are brain states of type $C$ but no consciousness, that is, it will be conceivable that (2) is only contingently true.

Chalmers argues that neither two-dimensionalism nor Kripke’s redescription strategy can help the friend of a posteriori physicalism here, but it is unclear to me that he is right to say this. To see why, let us grant, with Chalmers, that the primary and secondary intensions of the concept of being conscious coincide, and let us focus instead on the concept of being in a brain state of type $C$.

According to Chalmers’ own principles, the concept of being in a brain state of type $C$ is a paradigm example of a physical concept. Because it is a physical concept, however, it is very plausible to suppose that it—like the concept of being $H_2O$, or the concept of having molecular K.E.—has two distinct intensions. Its primary intension reflects the way in which reference to brain states of type $C$ was initially fixed; presumably, this was via a cerebroscope or some similar measuring device. The secondary intension of the concept of being in a brain state of type $C$, on the other hand, picks out brain states of type $C$ in virtue of their essential properties. Our first observation, then, is that the primary and secondary intensions of the concept of being in a brain state of type $C$ are distinct.

It should also be granted, however, that the primary intension of the concept of consciousness is distinct from the primary intension of the concept of being in a brain state of type $C$. For on the assumption that a posteriori
physicalism is true it follows that there is no logical connection between physical nature and psychological nature. So in particular, on the assumption that a posteriori physicalism is true it follows that there is no logical connection between the concept of being conscious and the concept of being in a brain state of type C. This is our second observation.

I have argued that the primary intension of the concept of being in a brain state of type C is distinct from its secondary intension, and from the primary intension of the concept of being conscious. It is therefore open to physicalists to say that a world which appears to contain brain states of type C but no consciousness is really a world in which the primary intension of the concept of consciousness is satisfied—i.e., is a world in which cerebroscopes give the same readings that they give when brain states of type C are triggered in the actual world—but in which neither the primary nor the secondary intension of the concept of consciousness is satisfied. That is, it is open to physicalists to claim that what is being imagined to be true when it is imagined that zombie worlds are possible is not (3):

\[(3) \quad x \text{ is in a brain state of type } C, \text{ and } x \text{ is not conscious.}\]

but is instead (4):

\[(4) \quad x \text{ is in a state that gives rise to readings of type } R \text{ on a cerebroscope, and } x \text{ is not conscious.}\]

(4) is the result of reinterpreting (2) according to the primary intension of the concept of being in a brain state of type C. But (4)'s being true does not constitute a counterexample to the claim that physicalism is true of the actual
world, since a world of which (4) is true need not be a world in which there are brain states of type C in the absence of consciousness. In short, Chalmers’ argument for the conceivability of zombie worlds is plausible only because he ignores the redesription strategy according to which the primary intension of the concept of being in a brain state of type C is at issue. Chalmers argues that since the primary intension and the secondary intension of the concept of being conscious coincide, there is no way to redescribe (2) in a way that does not reflect badly on the claim that it is necessarily true. But if we simply ignore the concept of being conscious and focus instead on the concept of being in a brain state of type C, the redesription strategy is applicable. Consequently, since it is always open to physicalists to redescribe worlds which seem physically indistinguishable from the actual world as worlds which only appear to be physically indistinguishable from the actual world, Chalmers’ argument for the conceivability of zombie worlds loses its force.

7. I can think of two responses to this argument. First, somebody might argue that it is an essential property of brain states of type C—or whatever physical property consciousness is said to be identical with, or supervene on—that they give rise to readings of type R on cerebroscopes. This objection is not to the point, however. For this issue is not whether brain states of type C necessitate readings of type R on cerebroscopes. Rather, the issue is whether there could be readings of type R in the absence of brain states of type C. So long as this is a possibility, the redesription argument remains unaffected.

Second, it might be objected that the claim that putative zombie worlds are in fact worlds which are physically distinguishable from the actual world
is implausible. Chalmers himself presses this objection.20 I focused on a simple conditional linking physical nature and psychological nature. But suppose we focus instead on a more complicated conditional linking physical nature and psychological nature, for example, \( \phi_{@} \supset \psi_{@} \) (where again \( \phi_{@} \) is the complex physical sentence that states the entire physical nature of the actual world, and \( \psi_{@} \) is the complex psychological sentence that states the entire psychological nature of the actual world). Then it might be argued that the redescription strategy described above is not applicable.

This objection has two parts.21 The first part alleges that, contrary to what I have argued, there is not a difference between the primary and the secondary intensions of the physical sentence \( \phi_{@} \). The second part alleges that if the primary and the secondary intensions of \( \phi_{@} \) diverge, then it follows that the physical nature of the actual world is essentially unknowable.

This is a powerful objection. Nonetheless, it seems to me that the basic point remains unaffected. As for the first part of the objection, since \( \phi_{@} \) is a physical sentence, it is very plausible to suppose that it has two distinct intensions. Whether we are in a position to spell out the primary intension of \( \phi_{@} \) is irrelevant; all that matters is that it has one. So long as we stick with the primary intension of \( \phi_{@} \) then it will seem possible that \( \phi_{@} \supset \psi_{@} \) is false. But this does not imply that \( \phi_{@} \supset \psi_{@} \) is false according to its secondary intension, and that is the intension that is relevant here.

The second objection is harder to rebut. I have been arguing that we can account for the apparent conceivability of zombie worlds without having to admit that such worlds are conceivable. And I suggested that we can do so


21 I am indebted to Alex Byrne here.
by redescribing the necessary conditionals which are claimed to be false—such as \( \phi_@ \supset \psi_@ \)—using the primary intensions of the physical concepts involved. The problem with this strategy is that it seems to have the unwanted consequence that we can only know about the physical nature of the actual world according to the primary intensions of physical concepts. But if essential properties are tied to the secondary intensions of such concepts—as, in effect, Kripke has argued—this seems to preclude our ever being in a position to have complete knowledge of the physical nature of the actual world. As Chalmers puts it, “it might be that for something qualify as an electron in a counterfactual world, it is not sufficient that it be causally related to other physical entities in the way that an electron [actually] is. Some hidden essence of electronhood might also be required...The essential nature of electrons...would then be hidden to physical theory, which characterizes electrons...only extrinsically.” (Chalmers 1996, 135)

Now, although it might seem somewhat odd, this sort of position is not without precedent. For there is a tradition in philosophy of thinking of the concepts and properties of physical theory as being dispositional in nature. And if this dispositional view is conjoined with the further view, (i), that physical theory discovers only dispositional properties, and (ii), that dispositional properties require non-dispositional properties as categorical bases, then we have a view according to which there are physical properties which lie outside the domain of physical theory.\(^{22}\)

\(^{22}\) See Blackburn (1992) for an articulation of the view that it is ‘dispositions, all the way down’. Armstrong (1968) endorses the idea that physical theory is concerned only with dispositional properties, and therefore concludes that physical theory does not tell us about the categorical bases of such properties. Stoljar (1998) has an interesting discussion of these issues.
Chalmers calls such inaccessible physical properties—misleadingly, in my opinion—*protophenomenal properties*. And his objection is that this sort of view collapses into a form of dualism which is no different from the view that he is arguing for. As he says, on this view “[a]fter ensuring that a world is identical to ours from the standpoint of our physical theories, God has to expend further effort to make that world identical to ours across the board. The dualism of ‘physical’ and ‘nonphysical’ properties is replaced on this view by a dualism of ‘accessible’ and ‘hidden’ physical properties, but the essential point remains.” (Chalmers 1996, 136) But does it? It does if we make the further assumption that two worlds which are indistinguishable from the standpoint of physical theory are physical duplicates. But since this is precisely what is at issue it can hardly be taken for granted. There is certainly no reason to suppose that physical theory must be capable of explaining *all* physical features of the actual world, and in fact we have seen that there is an argument for the conclusion that physical theory—due to the fact that it invokes only dispositional properties and concepts—cannot do so. Although these remarks are not demonstrative, it nonetheless seems to me that the idea that there are aspects of physical nature that are not accessible to physical theory is defensible, as is the idea that there could be a world which is indistinguishable from the actual world with respect to physical theory, but which nonetheless is not a physical duplicate of the actual world.

A satisfactory resolution of this issue would necessarily involve a thorough sorting-out of what is meant by the expression ‘physical property’. If ‘physical property’ is simply taken to mean ‘a property describable by physical theory’, then such protophenomenal properties will not be physical properties and this view will, as Chalmers alleges, collapse into a kind of dualism. On the other hand, if it is plausible to suppose that there could be
physical properties which cannot be described using the resources of physical theory then this view will be a physicalist view.

Although I cannot resolve this issue here, it seems to me that there is a way to account for the conceivable truth of the statement ‘there are zombie worlds’ without having to acknowledge the genuine conceivability of such worlds. I have argued that it is possible to do this if we employ a version of Kripke’s redescription strategy. This strategy was initially introduced in the context of necessary identity claims, but I think that it can be deployed with equal effectiveness in the case of entailment claims of the sort to which I have argued physicalism is committed.23

8. I have been arguing that, on reflection, zombie worlds are not conceivable. Why then do so many claim to find such worlds conceivable? I think the answer is not hard to find. To conceive of a world which is physically identical with the actual world might seem easy enough to do. And

23 Chalmers would presumably disagree. He writes:

It is crucial that the argument as I have put it does not turn on questions of identity but of supervenience. The form of the argument is not, ‘One can imagine physical state P without consciousness, therefore consciousness is not physical state P.’ The form of the argument is rather, ‘One can imagine all the physical facts holding without the facts about consciousness holding, so the physical facts do not exhaust all the facts.’ This is an entirely different sort of argument. In general, modal arguments for dualism that are cast in terms of identity are less conclusive than modal arguments cast in terms of supervenience.[.] (Chalmers 1997, 131)

I find these remarks puzzling. First, I do not see that the form of modal arguments for dualism that are cast in terms of identity are importantly different from those cast in terms of supervenience, as both appeal to the conceivability of situations where physical nature and psychological nature come apart. Second, it seems to me that modal arguments for dualism that are cast in terms of identity are just as effective as those cast in terms of supervenience. For if modal arguments for dualism that are cast in terms of supervenience are compelling, then corresponding arguments cast in terms of identity will also be compelling. That is, if, for example, we have an argument for the conclusion that the property of being F is does not supervene on the property of being G then we will also have an argument for the conclusion that the property of being F is not identical with the property of being G.
to conceive of a creature which lacks consciousness would appear to be fairly straightforward as well: simply imagine a creature for which ‘all is dark inside’. So to conceive of a zombie world I need only conceive of a world which is physically identical with the actual world, and which contains creatures for which all is dark inside. Does it follow that I have thereby succeeded in conceiving of a zombie world, or in showing that such worlds are conceivable? I have certainly succeeded in saying what a world would have to be like in order for it to be a zombie world; but to do that, I maintain, is not to show that such a world is conceivable. For it does not follow from the fact that an A-world is conceivable, and that a B-world is conceivable, that an (A&B)-world is conceivable. For example, I can conceive of a world in which all objects are round; and I can conceive of a world in which all objects are square; but I cannot conceive of a world in which all objects are both round and square (which is not to say that I cannot imagine what a world would have to be like in order for it to be a world in which all objects are both round and square).

What makes it particularly difficult to determine whether a statement is genuinely conceivable is that it is unclear what background conditions should be held fixed.\textsuperscript{24} Suppose somebody claims that it is conceivable that the moon should be made of green cheese. Is this really conceivable? Holding everything except for the fact that the moon is made of green cheese fixed, it might be thought that it is. But the fact is that we cannot hold everything fixed if our attempt to conceive that the moon should be made of green cheese is to be a serious one. In this I agree with Peter van Inwagen. Van Inwagen remarks that

\textsuperscript{24} What follows owes much to Sneddon (1972).
anyone who thinks he can imagine that the moon is made of [green] cheese has a very sluggish imagination: the active imagination demands a pasture for the antecedently necessary thousands of thousands of millions of cows, demands a way to preserve a piece of cheese in broiling heat, freezing cold, and vacuum for thousands of millions of years, demands some off-stage *machina* to protect a piece of cheese thousands of miles across from gravitational compression into non-cheese...But any *serious* attempt to imagine the moon being made of green cheese must, like the unimaginable object itself, soon collapse under its own weight. (van Inwagen, 1979, 671-2)

The example is somewhat fanciful, but van Inwagen's point is not. Depending on what background assumptions are held fixed, some statements will seem initially conceivable. But as soon as we begin to consider what follows from the supposed conceivability of a statement $E$, it will often become unclear that what we are being asked to conceive of is genuinely conceivable. The case of the moon being made of green cheese is one such example; the case of zombie worlds is plausibly another. So while somebody might object that an inability to conceive of or imagine zombie worlds indicates a defect in imagination, it seems to me that such an inability indicates instead an imagination which is overly active.\(^{25}\)

8. To conclude, I have tried to show that Chalmers' argument from the conceivability of zombie worlds to the conclusion that physicalism is false fails. I argued, first, that Chalmers equivocates on the crucial notion of

\(^{25}\) Of course, somebody might object that it is van Inwagen's imagination that is sluggish. For can we not imagine an enormous and lush pasture stretching off into the distance as far as the eye can see, on which roam cows which require only minimal amounts of grazing space and which produce large amounts of milk extremely quickly? And couldn't we imagine a moon that is made of a very hard Parmesan-like cheese, so hard that it is impervious to broiling heat and freezing cold? In response I can only say that, like van Inwagen, my imagination begins to give out at this point.
conceivability. In particular, I argued that the apparent conceivability of zombie worlds does not entail that such worlds are metaphysically possible. I then argued that there is a way to account for the apparent conceivability of zombie worlds without having to embrace the claim that zombie worlds are conceivable. As I said at the outset, it is a matter of considerable debate whether conceivability arguments in the philosophy of mind are capable of doing the work they are asked to do. The considerations of the present paper suggest that they are not.26

26 I would like to thank Alex Byrne, Robert Stalnaker, and Judith Jarvis Thomson for very helpful criticism and advice.
References


Byrne, A. “Cosmic Hermeneutics.” MS.


