# Conceivability, Possibility and the Mind-Body Problem

KATALIN BALOG

[Published in *The Philosophical Review*, Vol. 108, No 4. (Oct. 1999), pp. 497-528.]

> The feeling of an unbridgeable gulf between consciousness and brain-process: how does it come about that this does not come into the considerations of our ordinary life? This idea of a difference in kind is accompanied by slight giddiness—which occurs when we are performing a piece of logical sleight-of-hand.
>
> – Wittgenstein, *Philosophical Investigations*, §412

I want to take on the question of what a class of arguments, usually called the Conceivability Arguments, have to say about the mind-body problem. These arguments have two different versions. In one version, considerations of conceivability are taken to support the claim that phenomenal consciousness is not identical to, realized by, or supervenient on physical properties (for example, Kripke 1972, Nagel 1974, Robinson 1993, White 1986, Jackson 1998, and Chalmers 1996). According to the other version, there is an explanatory gap between phenomenal and physical levels of description that does not exist with respect to other higher-level descriptions and that may have metaphysical ramifications.[1] My claim is that these arguments do not succeed in establishing their conclusions. That is because (and I take this to be the primary lesson of the Conceivability Arguments) what they reveal does not have to do with phenomenal consciousness *itself*, but rather with the nature of *phenomenal concepts.*

---

I would like to thank John Biro, Ned Block, David Chalmers, Jennifer Church, Jerry Fodor, Gary Gates, Tamar Gendler, Joe Levine, Brian Loar, Barry Loewer, Colin McGinn, Brian McLaughlin, Karen Neander, Jesse Prinz, Georges Rey, Howard Robinson, Zoltán Szabó, Gene Witmer, and four anonymous referees for helpful comments and conversation.

[1]This argument is formulated by Joseph Levine (1998), although he does not endorse the conclusion.

In what follows, I will focus on the most elaborate and sophisticated version of the Conceivability Argument for dualism. First I provide a general exposition of the structure of Conceivability Arguments, then I proceed to describe in greater detail Frank Jackson's and David Chalmers's new Conceivability Argument. Finally I construct a reductio that at the same time reveals where the arguments went wrong.

## 1 Introduction

Phenomenal consciousness—the *what it's like*[2] feature of experience—can appear to a scientifically inclined philosopher to be deeply mysterious. It is difficult to conceive of how the swirl of atoms in the void, the oscillation of field values, or anything physical can add up to the smells, tastes, feelings, and so forth that constitute our phenomenal experience.

The most important argument for the claim that there is no place for phenomenal consciousness in a completely physical reality relies on considerations of *conceivability*. The argument, which goes back at least to Descartes (Sixth Meditation, in Cottingham, Stoothoff, and Murdoch 1984, 2:50-63), begins with the premise that we can conceive of *any* physical or functional facts obtaining without there being any phenomenal experience at all.[3] This is sometimes expressed by saying that zombies (that is, beings that are our physical and functional duplicates, but possess no phenomenal experiences) are *conceivable*.[4] From this assertion of conceivability it is inferred that zombies are genuinely possible. And this conclusion is incompatible with physicalism as that doctrine is usually understood.

---

[2]The expression is coined by Thomas Nagel (1974).

[3]Sometimes it is argued that the opposite is also conceivable, that is, that it is conceivable that mental facts, especially experiences, occur without any physical or functional facts occurring (see Descartes, Sixth Meditation). It is not necessary, however, for the arguments under consideration, that conceivability go both ways.

[4]I will use the terms 'experience', 'phenomenally conscious state', and 'phenomenal state' interchangeably. The phenomenal aspect of a mental state is the same as its experiential character, or, in Nagel's (1974) words, its 'what it is like' feature.

The claim that zombies are *conceivable* does not have to do with our powers of imagination, or our *psychological* constitution in general, but rather with the nature of physical and phenomenal *concepts.* The relevant notion of conceivability is this:

(Con) A statement S is conceivable if it is consistent with the totality of conceptual truths, that is, if –S is not a conceptual truth.

Conceptual truths (or analytic truths) are truths in virtue of meaning.[5] It is usually assumed that if S is conceivable then it is knowable *a priori*—at least in principle, by an ideal logician—that S is conceivable. In other words, someone who can entertain the thought that S can come to know whether S is conceivable without empirical investigation. Further, failure to detect a priori any contradiction in S is taken as a defeasible reason to hold that S is conceivable. It is defeasible since further a priori reasoning may lead one to see that S is inconsistent with analyticities after all.[6]

To support the premise that zombies are conceivable, it is claimed that there is no contradiction, detectable a priori, in describing a possible world as being physically exactly like our world, yet containing no experiences. Some philosophers have denied this: they claim that our *concepts* of various kinds of phenomenal states (for example, pain) are physical, functional, or behavioral concepts (Lewis 1966, Ryle 1949, White 1986). For example, a crude functionalist account of the concept pain is that it is the concept *internal state produced by stimuli associated with harm and typically causing aversive be-*

[5]The nature of concepts, what determines whether a statement or thought is true in virtue of meaning, and even whether there are any conceptual truths at all are vexed and disputed matters; see Fodor 1997. Since the proponents of Conceivability Arguments rely on the notion of conceptual truth, I will as well.

[6]The claim that whether S is conceivable is always knowable a priori is not quite correct, since logical consistency is not effectively decidable and, if the underlying logic is higher order, not even effectively axiomatizable. But this observation has no effect on the Conceivability Arguments.

*havior*. Of course, if it is analytic that an internal state satisfying a certain functional specification is pain, then zombies are inconceivable.[7]

It seems to me that behaviorist and functionalist analyses of phenomenal concepts are quite implausible. When I think (the same for you, I submit) *I am in pain* I am not thinking that I am behaving or disposed to behave in some way, or that I am occupying some particular neurophysiological state or functional state. Of course, this is not to say that the property of being in pain is not a physical or functional property, but rather that the concept *pain* is not a functional or physical concept. Whatever the ultimate nature of phenomenal experience, when I judge that I am having an experience of a particular sort on the basis of actually having that experience, the concept I invoke is not a behavioral, physical, or functional concept. Rather, it seems to be a concept that I apply directly and spontaneously to the experience.[8]

There is another line of reasoning that can be seen as aiming to show that zombie-worlds are inconceivable. I have in mind Wittgenstein=s *private language argument*.[9] The argument relies on certain a priori considerations concerning the nature of meaning. The basic idea is that first-person direct uses of a phenomenal concept presuppose that the concept has links with publicly observable behavior (or other physical phenomena) that provide criteria for third person uses. These criterial connections are alleged to preclude zombie-worlds. It would be well beyond the scope of this paper to evaluate this argument. But in any case, in the following I want to grant to the proponent of the Conceivability Arguments as much as possible. So I will grant that there is nothing in our

[7]Another view contrary to the conceivability of zombies relies on the claim that, while concepts of kinds of experience—for example, pain, nausea, etc.—do not have a functionalist analysis, the concept *conscious experience* does. For example, Shoemaker (1981) holds that zombies are inconceivable, but that inverted qualia are conceivable (indeed, possible). This is an interesting view, but in itself is not enough to block the Conceivability Arguments.

[8]Loar (1997) characterizes phenomenal concepts as 'direct recognitional' concepts.

[9]Wittgenstein 1953, §§ 207-384. The argument is usually invoked in the discussion of 'other minds.' But of course the question of whether another being has a mind is just the question of whether she is a zombie.

concept of consciousness that would allow us to rule out a priori the existence of zombies: zombies are conceivable. I do not want the defense of physicalism to depend on either the private language argument or such a contentious semantic doctrine as analytic functionalism or analytic behaviorism about qualia.

The conceivability of zombies, however, is used to support the claim that zombies are *genuinely metaphysically possible.* This is a powerful result. If it is correct, and if, as I will assume throughout this paper, there are phenomenal facts, then physicalism is false. For it would mean that the totality of physical facts obtaining in our world, including nomological and causal facts, does not *necessitate* the phenomenal facts that obtain in our world.[10]

But there is an obvious objection. On the face of it, the mere fact that a state of affairs is conceptually possible does not entail that it is metaphysically possible. The mere fact that it is conceptually possible for F to exist without its being G does not entail that it is metaphysically possible for F to exist without being G. For example, it is conceptually possible (at least it was before the eighteenth century) that water is not $H_2O$, but it is not really metaphysically possible for water not to be $H_2O$, since water *is* $H_2O$, and we know from Kripke's (1972) work that identities, where the terms of identity are rigid designators, are necessary. But during the last three decades the relationship between conceptual possibility and metaphysical possibility has been greatly clarified (again, especially by the work of Kripke (1972)) so as to take these objections into account.

This has led to a revival of interest in Conceivability Arguments, and sophisticated versions of these arguments have been developed by Kripke (1972, 144-55), Nagel (1974, 435-50), White (1986, 333-68), Robinson (1993), Jackson (1982, 1993 and 1998, chaps. 2 and 3), Chalmers (1996, esp. 56-123), Levine (1998, 449-80), and others. Like their predecessors, these arguments rely on there being a link between conceivability and

[10]There is a form of this argument that does not aim at a metaphysical conclusion but merely at an epistemic one. In this form the argument aims to establish that there are features of conscious states that will forever elude scientific explanation. The position originates with Levine (1983, 1993).

metaphysical possibility, but in the formulation of this link they now take into account that conceivability does not *always* imply possibility. The proponents of these new Conceivability Arguments claim that while the conceivability of water's not being $H_2O$ fails to imply that it is *metaphysically* possible for water not to be $H_2O$, the conceivability of a zombie-world *does* imply that a zombie-world is a genuine possibility.[11]

As we will see, the link between conceivability and possibility invoked by Conceivability Arguments entails that all modal facts are ultimately reducible to facts about what is conceivable and ordinary empirical facts (including laws) that play a role in fixing the references of our concepts. In this way the link provides a very attractive picture of the metaphysics and epistemology of possibility. In this picture the *truth makers* of modal claims are not a realm of possible worlds, but rather facts about our concepts and ordinary empirical facts. And modal truths are knowable by a combination of a priori reflection on our concepts and empirical investigation. In fact the promise of this account may be the strongest reason for accepting some form of the conceivability-possibility link.

However, my aim here is to consider the new Conceivability Arguments due to Frank Jackson and David Chalmers and show that the very principle connecting conceivability and possibility they rely on is mistaken. While their arguments are my particular focus, my criticisms extend to the other Conceivability Arguments as well, since I will be attacking the link between conceivability and metaphysical possibility that they all presuppose. These arguments are all refutable by a master argument that I call 'the Zombie Refutation.'[12] The reason they fail has to do with the very nature of phenomenal concepts that gives rise to the conceivability of zombies. Because of the special nature of these concepts, the principle that links conceivability and possibility turns out to be self-refuting. Thus, the zombies that antiphysicalists think possible in the end undermine the

[11]Of course, they will argue that the difference is between *kinds* of statement. The claim is, as it will soon be clear, that there is a kind of statement for which conceivability implies possibility. The statement that a zombie-world exists is supposed to fall under this kind.

arguments that allege to establish their possibility—a fitting revenge. While these considerations fall short of establishing the truth of physicalism, they go a long way toward defending it from some of the most influential arguments against it. Although I agree with Jackson and Chalmers that there is something puzzling about consciousness, I do not think that the puzzle adds up to a refutation of physicalism.

## 2 The Argument

Jackson's and Chalmers's arguments are similar. Their definitions of physicalism are almost identical, as are the semantical frameworks in which they formulate their arguments. Although they employ slightly different formulations of the crucial premise linking conceivability and possibility, for present purposes I will assume that they employ the same one since it can be shown that Chalmers's premise entails Jackson's.[13] I will be mainly following Jackson's exposition, but my reconstruction of the argument is meant to be attributed to both of them.

### 2.1 Preliminaries

One caveat: Whereas Chalmers (1996) eagerly embraces the dualist conclusion of the argument, Jackson (1993 and 1998) has a more cautious attitude. He himself presents the argument as a challenge for the physicalist rather than a straight refutation of physicalism, and recently seems to reject its conclusion. But on plausible assumptions, shared by Jackson (1982), it can be easily turned into a refutation. And this is how I will treat it.

In a nutshell, the argument is the following: Physicalism requires that a phenomenal statement, like 'Frank is experiencing a yellow sensation', must, if true, be necessitated by truths expressed in the language of physics. Jackson and Chalmers argue that this necessitation must itself be a priori and that such a priori truths must be grounded in the nature of phenomenal and physical concepts. However, phenomenal concepts do not sup-

[12]For the extension of the refutation to other Conceivability Arguments see Balog 1998.

[13]See Balog 1998.

port such a prioricities. It follows, assuming that there are phenomenal truths, that physicalism is false. Let us now look at the argument a little more closely.

**Physicalism**

Jackson observes that physicalism, at a minimum, requires a commitment that

> (P) Any world which is a *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world.[14]

Two worlds are physical duplicates if and only if they agree on all the true statements expressed in the language of physics.[15] A *minimal* physical duplicate of our world is what we would get if we used the physical nature of our world (including, of course, the laws) as the *sole ingredient* in making a world (see Jackson 1993, 28); so, a *minimal* physical duplicate of our world is, by definition, a physicalistic world.[16]

Jackson intends this to capture the idea that there is nothing over and above the physical stuff in our world.[17] He suggests that his formulation of physicalism (P) is equivalent to the claim that every truth about our world, be it physical, chemical, biological,

[14]Chalmers gives essentially the same definition (1996, 41-42). Their formulation is similar to Lewis's definition in his 1983. For expository reasons, I will stick with Jackson's formulation throughout the paper.

[15]The exact content of *physicalism* depends on exactly how physics is understood. Current physics is almost certainly not exactly true or complete and we have no idea how to characterize future physics. But it suffices for Jackson's (and my) purposes to assume that the language of physics includes no mentalistic (that is, phenomenal or intentional) vocabulary (see Papineau 1993, 29-32).

[16]Unless, of course, there are nonphysical entities or properties that are connected, with metaphysical necessity, to physical entities or properties. But this is a complication I wish to ignore for now.

[17]The definition also captures the intuition that physicalism is a contingent doctrine, that is, that physicalism is true in some worlds, but false in others. For example, a *minimal* physical duplicate of a world containing ghosts will not be a duplicate simpliciter of the ghost-world, since the ghosts will be missing in the minimal physical duplicate world. On the other hand, a minimal physical duplicate of a physicalistic world will be a duplicate simpliciter of that world.

cal, psychological, etc., is necessitated by a statement of physics that gives the full physical description of the world, and is true in all and only the minimal physical duplicates of our world. For the purposes of this paper, I will use the following definition:

(E) For any true statement T, □(K -> T),

where K is a very long conjunction, expressed in the language of physics,[18] giving the complete physical truth (including truths about the laws of physics) about our world.[19]

**Conceptual Explanation**

According to Jackson and Chalmers, the necessities 'K -> T' (where T is a truth) cannot be brute facts; they need explaining.[20] Jackson maintains that if T is, for example, a psychological statement, then analytical functionalism has a story to tell about why the statement is necessary. As he puts it:

[18]This definition is not strictly equivalent to (P). Statements that make reference to special kinds of property—to put it crudely, global properties—are not necessitated by the full fundamental description of the world K; they are only necessitated by the conjunction of K with the statement that K *is* the full fundamental description of the world. However, consciousness, among many others, is arguably not such a property. Whether I am in pain is a positive fact that does not seem to depend on any other fact except local facts about me. So, for the purposes of this paper, this issue can be safely ignored.

[19]More formally, the definition is: (E) (Y)(Y -> □(K -> Y)), where Y is a sentential substitutional quantifier.

[20]The point is, actually, not just that (E) (and (P)) needs an explanation; it is that there *are* explanations of (E) that are actually incompatible with physicalism. For example, (E) could be true in virtue of some strange set of 'quizzical' properties that underlie both physical and nonphysical property instantiations (see Witmer 1997, 137). But this issue will not affect the main argument. Even if (E) is not sufficient, it is still necessary to make physicalism true, so if *it* can be refuted by Jackson and Chalmers, physicalism is refuted. (And, I might add, if *it* cannot be refuted, then we have no reason to doubt physicalism, since Jackson and Chalmers provide no further reasons against physicalism.)

> it is the very business of conceptual analysis to explain how matters framed in terms of one set of terms and concepts can make true matters framed in a different set of terms and concepts. (1993, 32)

Jackson's view is that in the absence of a conceptual story of how the purely physical makes the psychological true, the entailment would remain an 'impenetrable mystery'. Both he and Chalmers think that the explanation has to be, in an appropriate sense, *conceptual.*[21]where K* is the full description of the world in the language of fundamental discourse, and T is any truth. (In the case of Berkelean idealism, for example, the fundamental discourse is mentalistic, and all the physical truths have to be a priori entailed by a complete mentalistic description of the world.) They argue that if physicalism is true, then 'K -> T' is not only metaphysically necessary, but it is also an a priori conceptual truth; that is, they argue that if physicalism is true, then all truths are a priori derivable from the full physical description of the world. I will call this the

> *A Priori Entailment Thesis*: If (E) is true, then, for any true T, statements of the form K -> T are conceptual truths.[22]

This is the key premise in Jackson's and Chalmers's argument against physicalism; it provides the crucial link between conceivability and possibility.

---

[21]This, of course, is not an arbitrary requirement for physicalism alone. Jackson is explicit that any metaphysical theory that makes a distinction between fundamental and nonfundamental properties—for example, Berkelean idealism, or Cartesian dualism—has to be able to produce, for any true T, appropriate derivations of the respective entailment claims K* -> T, where K* is the full description of the world in the language of fundamental discourse, and T is any truth. (In the case of Berkelean idealism, for example, the fundamental discourse is mentalistic, and all the physical truths have to be a priori entailed by a complete mentalistic description of the world.)

[22]More formally: (Y)(Y -> □(K -> Y)) -> (Y)(Y -> 'K -> Y' is a priori knowable), where Y is a sentential substitutional quantifier.

Why think that the *A Priori Entailment Thesis* is true? Jackson provides the following considerations. First of all, he claims that many truths conform to it, and there is no reason to suppose that some will not; also, it is immune to the criticism we made earlier with respect to the naive conceivability-possibility principle. Although it is conceivable *simpliciter* that water is not $H_2O$, it is not conceivable *consistent* with the full physical description of the world. Building on Kripke's argument (1972, 140-162), Jackson observes that, arguably, in all *bona fide* cases of identity statements where the denial of the identity statement is conceivable (for example, 'water is not $H_2O$'), there are contingent truths, which, together with conceptual truths involving the terms in question (here, the terms 'water' and '$H_2O$'), *entail* the identity statement.

For example, on the assumption, roughly, that $H_2O$ is the unique thing that plays the water-role, the statement that water is not $H_2O$ is not conceivable, since it is a conceptual truth that the unique thing that plays the water-role *is* water.[23] Jackson generalizes this observation and claims that the denial of all *bona fide* true statements, in conjunction with the *full fundamental truth* about the universe, is inconceivable. His idea is that the full fundamental description of the universe always provides enough background information to fix the reference of any concept in terms of fundamental concepts, and so it is always possible to derive any true statement from it.

Let=s look at the example involving water and $H_2O$ in some detail. Suppose water covers 60% of the surface of the Earth. Then, according to Jackson, it can be shown that the statement

(W) K -> water covers 60% of the surface of the Earth

is a priori. Let's see how. Jackson claims that something like the following is an a priori truth

[23]That it is a conceptual truth follows from Jackson's semantics. Here we do not have space to give it the full treatment it deserves.

(i) Water is the clear, odorless, etc...liquid around here that fills the oceans and lakes, etc.

It follows a priori from (i) that

(ii) $H_2O$ is the clear, odorless, etc. ...liquid around here that fills the oceans and lakes, etc. -> Water is $H_2O$.

But it is also a priori true that

(iii) (Water is $H_2O$) -> (($H_2O$ covers 60% of the surface of the Earth) -> (Water covers 60% of the surface of the Earth)).

From (ii) and (iii) we get

(iv) $H_2O$ is the clear, odorless, etc....liquid around here that fills the oceans and lakes, etc. -> ($H_2O$ covers 60% of the surface of the Earth -> Water covers 60% of the surface of the Earth).

But this is equivalent to

(v) ($H_2O$ is the clear, odorless, etc....liquid around here that fills the oceans and lakes, etc. & $H_2O$ covers 60% of the surface of the Earth) -> Water covers 60% of the surface of the Earth.

If this derivation is correct, we have shown that the statement

H -> Water covers 60% of the surface of the Earth,

where H is a conjunction of contingent statements about $H_2O$,[24] is a priori.[25] Since, according to Jackson, these contingent statements about $H_2O$ are similarly a priori derivable, perhaps through some intermediary steps, from contingent truths of microphysics, we have shown that

(W) K -> Water covers 60% of the surface of the Earth

is knowable a priori. Jackson thinks that most true statements[26] can be similarly shown to be a priori entailed by the full physical description of the world.

A further motivation for the *A Priori Entailment Thesis* is that it is a very powerful explanatory claim. Modal claims of the form

□(K -> T)

might seem metaphysically and epistemically mysterious. If correct, the *A Priori Entailment Thesis* would explain these necessities in terms of conceptual truths, and it would explain metaphysical necessity in general in terms of conceptual necessities and contingent truths, since, according to it, the statement

K -> M,

[24]H is a conjunction of statements like, for example, '$H_2O$ is clear, odorless liquid.'

[25]Of course, the derivation, as it stands, is incomplete. To complete it, we would have to have the requisite conceptual truths that link the concepts 'Earth', 'surface', 'clear', 'odorless', etc. to terms of lower-level discourse, and, ultimately, microphysics. According to Jackson, it is clear that such conceptual truths exist.

[26]With the exception of phenomenal statements. But, of course, he thinks that even those would be entailed a priori by the *full fundamental description* of the world.

where K is the full fundamental description of the world, and M is any metaphysical truth, is a conceptual truth.[27] This means that any metaphysically necessary truth M can be conceptually derived from K, the totality of contingent fundamental truths. This account also provides an epistemology for modality.

To recap, the support Jackson (and Chalmers) offer for the *A Priori Entailment Thesis* is this: there are good explanatory motivations for it; and in fact many putative necessities of the form K -> T do demonstrably conform to the *A Priori Entailment Thesis*. The claim is that there is no reason to suppose that there are exceptions to it.[28] The main goal of this paper is to give such reasons. I will show that, contrary to Jackson, there *are* exceptions to the *A Priori Entailment Thesis*.

### 2.2 The Argument

If the *A Priori Entailment Thesis* is true, the physicalist faces trouble *vis-à-vis* fitting psychological, and especially phenomenal, properties into the physical world. The reason is that there are no suitable conceptual analyses of phenomenal concepts for the relevant supervenience claim

K -> x feels pain (or any other statement expressing a phenomenal proposition)

to be a priori.

---

[27]This statement has to be qualified somewhat. Metaphysical claims like 'universals are prior to tropes' or 'fundamental properties are categorical' apparently can be denied without conceptual incoherence. Since it is implausible that the full fundamental description of the world will settle their truth, they are not going to be a priori derivable from that description. It is an interesting question how, on Jackson's view statements of this type should be handled; but we cannot discuss this here.

[28]Jackson and Chalmers also supply a much more sophisticated and elaborate argument for the *A Priori Entailment Thesis*, based on the so-called two-dimensional semantic framework. They seem to suggest that two-dimensional semantics, together with uncontroversial claims, entails the *A Priori Entailment Thesis*. However, it can be shown that the argument is question-begging (see Balog 1998).

The derivation of 'K -> Water covers 60% of the surface of the Earth' depended on the conceptual truth 'Water is the clear, odorless, etc. ... liquid'. The availability of such conceptual truths is essential to the kind of derivation we are considering, since the derivation works by finding a contingent statement linking the description to a term of a lower-level theory, and ultimately to a term of microphysics. Now consider the statement

K -> x feels pain.

To derive 'x feels pain' a priori from K, there must be some conceptual truth connecting 'pain' with a *nonphenomenal* description such that satisfaction of the description is a priori sufficient for 'feels pain'. But, arguably, there are *no* such conceptual truths.[29] For any

[29]One might think that, on the model of our derivation of 'Water covers 60% of the surface of the Earth', we can derive a priori, for example, 'x had pain' from contingent truths, if we allow just *any* contingent truths to figure in the derivations. For imagine the following argument:

(a) x has C-fibre firing (contingent empirical truth).
(b) Pain is the originating cause of pain-behavior (contingent empirical truth).
(c) C-fibre firing is the originating cause of pain-behavior (contingent empirical truth).

From (b) and (c) we get

(d) Pain is C-fibre firing.

From (a) and (d) we get

(e) x has pain.

This derivation uses only contingent empirical truths and conceptual truths; it shows that

P -> x has pain,

where P is a conjunction of contingent facts of neurophysiology and psychology, is knowable a priori. The problem with this derivation, however, is that one of the conjuncts in P, premise (b), is not *itself* a priori derivable from K; and if physicalism is true, according to Jackson, (b) could be true only if it were so derivable.

such nonphenomenal description we can *conceive* of its being satisfied without anyone feeling pain. 'Pain' is, as Loar (1990) calls it, a direct recognitional term; we do not apply the term, at least in our own case,[30] on the basis of any evidence, sensory, behavioral, or physical, distinct from what the term picks out, that is, distinct from the experience itself. 'Pain' refers to pain directly, or rather, via an essential feature of it, say, painfulness.[31] But it follows from the *A Priori Entailment Thesis* that if 'x feels pain' cannot be derived *a priori* from K, then

□(K -> x feels pain)

is false, and so if 'x feels pain' is true[32], then physicalism is false.[33] To put it more formally:

(1) If physicalism is true, then for any true T, statements of the form K -> T are conceptual truths.

(2) There are some true statements Q to the effect that phenomenal conscious experience occurs (eliminativism about phenomenal experience is false).

---

[30]And for phenomenal concepts, application to others is arguably derivative on the first-person use of the term (see Loar 1990).

[31]See Loar 1990 and Sturgeon 1994 for a discussion of this. In fact, on a Kripkean direct reference theory, this also applies to proper names, demonstratives, natural kind terms, etc. The point is that on *Jackson* and *Chalmers=s* view, this feature is unique to phenomenal concepts.

[32]Another way to block the argument is to deny that there are phenomenal states; see, for example, Rey 1988. But, again, I put this objection to the Conceivability Argument aside, since I do not want a refutation of it to depend on such a controversial claim.

[33]As we have already pointed it out, Jackson is not explicit about this. But in his 1982 he provides the tools to generate trouble for the physicalist from the *A Priori Entailment Thesis*. In that paper Jackson maintained that Mary is not able to deduce, even from the full physical description of the world, that a certain phenomenal experience, for example,

(3) If Q is a phenomenal statement, then 'K -> Q' is not a conceptual truth.

So

(4) Physicalism is false.

This is a really remarkable result. One might wonder, however, about its conclusion. What would the world be like if dualism were true? First, the dualist has to account for why psychophysical correlations occur even though phenomenal states do not metaphysically supervene on the physical. Nomological correlations have to be posited to hold the two realms together; but that leads to an ontology with a multitude of fundamental laws connecting complex physical states with apparently simple phenomenal states. These fundamental laws would be different from any laws of nature we know from science. Second, a dualist would either have to deny the causal closure of physics, countenance implausible causal overdetermination, or accept epiphenomenalism for phenomenal states.[34] None of these options is very attractive, however. Chalmers seems to prefer epiphenomenalism, but that would make it completely mysterious how we know about our own phenomenal states.[35] Third, although the new Conceivability Arguments rely solely on the conceivability of worlds exactly like ours physically, but lacking any phenomenal properties instantiated, and not on the converse, that is, the conceivability of worlds exactly like ours phenomenally, but lacking in any physical properties instantiated, it appears that an advocate of the Conceivability Arguments would have to condone

---

red phenomenal experience, occurs. There are indications that he might have changed his mind on exactly this issue (see his 1996, 142.)

[34]The most important formulations of the argument that shows this are by Papineau (1993), Loewer (1995), and Witmer (1997). They intend this as an argument *for* physicalism.

[35]Chalmers says that a person is *acquainted* with her phenomenal states and that this relation is not a causal one. But this seems to just put a label on the mystery.

the existence of purely phenomenal worlds.[36] It is barely intelligible what a world like that would be like. All these considerations make dualism very implausible. Fortunately, it can actually be shown that the arguments for dualism we have been considering are unsound.

## 3 Zombies Deceived

I now introduce the Zombie Refutation. This argument will show that the Conceivability Argument as formulated by Jackson and Chalmers is self-undermining, that is, that with the addition of some plausible further premises we can derive a contradiction from it. Suppose that Jackson's argument is sound. Its conclusion, that physical facts do not necessitate phenomenal facts, would then be true. And it would follow that there is a possible world that is exactly like our world physically, but in which no phenomenal, or other, nonphysical, facts obtain.[37] Let me emphasize: I make this assumption only for the sake of a reductio. Of course, if physicalism is true, as I think it is, then such a world is impossible. But my strategy is to show that the very assumption that there is such a world undermines the argument that leads to positing the existence of such a world.

In the world we are imagining there exists a zombie-Jackson, physically just like Jackson, but not the subject of any phenomenal states. Professor zombie-Jackson appears to give a series of lectures in zombie-Oxford (as Jackson did in Oxford) arguing for the *A Priori Entailment Thesis*. What are we to make of his words?

First of all, plausibly, zombie-Jackson will have intentional states. When he talks, his words are not mere meaningless sounds. I will argue that it is plausible to assume that zombie-Jackson has intentional states even if he lacks phenomenal states. Moreover, I will argue that it is plausible to assume that zombie-Jackson's intentional states will be identical with Jackson's intentional states except for intentional states that, in Jackson, involve phenomenal concepts. Those of zombie-Jackson's intentional states that, in Jack-

[36]Descartes actually did in the *Meditations*.

[37]In fact, this world would be a minimal physical duplicate of our world.

son, involve phenomenal concepts will refer to states of affairs present in zombie-Jackson's world. On this view, zombie-Jackson's argument will be just as meaningful as Jackson's, though not quite *identical* to it. Although the argument is *word by word* identical to Jackson's argument, some of the words (those that express phenomenal concepts in Jackson's language) have different meanings in Jackson's and zombie-Jackson's mouths. I mark these words with a '$^{+}$'. 'Pain$^{+}$', for example, stands for a term of zombie-Jackson that corresponds to Jackson's term 'pain'. They will use the same words to express different concepts; whereas Jackson's concept is phenomenal, zombie-Jackson's concept, by assumption, is not. We will come back to the exact nature of the difference shortly.

Zombie-Jackson's argument will go like this:

(1*) If physicalism is true, then for any true T, statements of the form K -> T are conceptual truths.

(2*) There are some true statements $Q^{+}$ to the effect that a phenomenal$^{+}$ state occurs (eliminativism about phenomenal$^{+}$ states is false).

(3*) If $Q^{+}$ is a phenomenal$^{+}$ statement , then 'K -> $Q^{+}$' is not a conceptual truth.

So

(4*) Physicalism is false.

My plan is the following: Given the assumptions I have made, I will argue that if a premise of Jackson's argument is true, the corresponding premise formulated by zombie-Jackson will be true as well. We know, however, that the dualist conclusion of zombie-Jackson's argument is false in the zombie-world. Remember, physicalism, [or anti-

physicalism], on the notion that is relevant here, is not a necessary doctrine; it is true about some worlds and false about others. The zombie-world, by stipulation, is a minimal physical duplicate of our world, so *its* minimal physical duplicates will be duplicates simpliciter *of it*. According to our formulation of physicalism, this is what it takes for a world to be physicalistic. Consequently, we know that zombie-Jackson's argument cannot be sound. Since, given that it is meaningful, it is clearly valid, one of its premises has to be false. It follows then that one of the premises of Jackson=s argument has to be false as well.

This plainly amounts to a reductio of Jackson's argument. It turns out that it is possible to derive a contradiction from Jackson's original premises, taken together with a few plausible additional assumptions. I will argue that these assumptions are indeed plausible, and that, given these assumptions, premises (1*)–(3*) follow from Jackson's original premises. But then the dualist conclusion of zombie-Jackson's argument follows as well, which contradicts the claim, also a consequence of Jackson's original argument, that the zombie-world exists and that it is physicalistic.

The fact that one can derive a contradiction from the original argument, together with the added premises, shows that one of the premises must be false. Since, I argue, [my added premises] are extremely plausible, the fault must lie with one of the premises of zombie-Jackson's argument (and, consequently, with the corresponding premise in Jackson's argument): the obvious candidate is the *A Priori Entailment Thesis*. While this does not necessarily mean that the dualist conclusion is false, it does mean that the argument used to establish it is not effective.

Let's now formulate these auxiliary assumptions more precisely. I am going to state them briefly right at the start; they will be discussed and defended in detail later, after the argument is given. Let me point out here that these assumptions do not by themselves imply physicalism (indeed, a dualist might very well accept them).

*Assumption 1:* Jackson and zombie-Jackson share most of their intentional states except those involving phenomenal concepts.

*Assumption 2*: Those concepts of zombie-Jackson that correspond to Jackson's phenomenal concepts will refer in the zombie to some (physical) state of the zombie.

*Assumption 3*: A prioricity for thoughts supervenes on the conceptual roles of their constituent (and related) concepts.

Let=s see how these assumptions, together with Jackson's premises (1)-(3) will suffice to derive zombie-Jackson's premises. As we said, zombie-Jackson, being Jackson's physical twin, offers an argument that is identical, word for word, to Jackson's argument. On *Assumption 1*, Jackson and zombie-Jackson mean the same by their words, except where phenomenal terms are involved. This means that premise (1*) in the zombie=s language expresses the *A Priori Entailment Thesis*, which, if true, is necessarily true, so if it was true in the actual world, it will be true in the zombie-world as well. On *Assumption 2*, we get premise (2*), that is, the claim that eliminativism about phenomenal$^+$ properties is false. Given *Assumption 3*, premise (3*) of zombie-Jackson's argument,

(3*) If $Q^+$ is a phenomenal$^+$ statement, then '$K \rightarrow Q^+$' is not a conceptual truth,

has as much claim to be true as premise (3) in Jackson's argument even though '$K \rightarrow Q^+$' has a different meaning from 'K -> Q'. The reason is that Jackson's phenomenal concepts and zombie-Jackson's phenomenal$^+$ concepts have *parallel* conceptual roles. $Q^+$, like Q, lacks conceptual links to physical, functional, and behavioral concepts sufficient to ground the a prioricity of '$K \rightarrow Q^+$'. On *Assumption 3*, a prioricity, or conceptual necessity, supervenes on the conceptual roles of the relevant concepts. That means that if 'K -> Q' is not derivable from conceptual truths given that Q lacks sufficient conceptual links

to physical, functional, and behavioral concepts, then neither is 'K -> $Q^+$' derivable from conceptual truths, since $Q^+$ will also lack the appropriate conceptual ties to physical, functional, and behavioral concepts.

I would like now to consider some objections. First, *Assumption 1*: One might object to it that zombies do not have intentional states at all. Presumably the reason would be that having phenomenal states is essential for having intentional states. In other words, one might object that because zombie-Jackson does not have phenomenal states, he does not really have *bona fide* intentional states either, and so cannot put forward any argument.[38] The most prominent exposition of this view is due to Searle (1992, chap. 7); he attempts to establish that consciousness[39] is necessary for intentionality. His argument is based on considerations about the inscrutability of reference originally formulated by Quine (1960, chap. 2). Searle puts his thesis in the following form:

> The notion of an unconscious mental state implies accessibility to consciousness. We have no notion of the unconscious except as that which is *potentially* conscious. (1992, 152; emphasis in original)

Contrary to Searle, a good case can be made that zombie-Jackson does have intentional states. Zombie-Jackson communicates with his colleagues: he answers questions, his utterances convey information, his actions are made intelligible by the assumption that he has beliefs and desires, etc. His cognitive organization seems to be essentially the same as Jackson's. On a more technical note: on all the extant theories of meaning, zom-

[38]The objection can be made more general by simply claiming that intentionality does not supervene on the physical. In this case, however, the argument for dualism based on qualia would already *presuppose* dualism about intentionality.

[39]Searle does not distinguish between phenomenal consciousness and the cognitive aspects of consciousness; he probably thinks that the two are metaphysically connected. In any case, I take him to say that all intentional states have to be at least potentially phenomenally conscious. This is the reading on which Searle's thesis causes a prima facie problem for my argument.

bie-Jackson will count as a thinker. On a Davidsonian interpretationist account, zombie-Jackson will have intentional states: he is just as interpretable as Jackson is. Similarly with other theories, like the informational account (for example, Dretske 1988), the causal-historical account (for example, Kripke 1972), the counterfactual account (for example, Fodor 1990), the teleosemantic account (Millikan 1989, Papineau 1993), etc. Zombie-Jackson's brain states (putting the problem of phenomenal versus phenomenal$^{+}$ states aside for the moment) carry the same information as Jackson's brain states; they have the same causal history linking them to entities in the world as Jackson's brain states do; the same counterfactuals hold about them as about Jackson=s brain states, etc. The only account on which zombie-Jackson will not count as a genuine thinker is the account on which phenomenal consciousness is essential for intentionality. Although the idea is not absurd, the argument for it does not seem to be very strong, and the contrary assumption seems far more intuitive. Moreover, the proponent of the Conceivability Arguments has to hold *both* that phenomenal consciousness is nonphysical *and* that it is essential for intentionality; but then we are owed an explanation of how causally inert, nonphysical properties can play a role in endowing mental symbols with meaning. In any case, since the thesis poses a serious challenge to my argument,[40] I would like to come back to it after I considered some other objections.

One might also object to *Assumption 2*, that is, the claim that the zombie's term 'pain$^{+}$' refers to a (physical) state of the zombie.[41] There are two ways in which *Assumption 2* could be false: first, if 'pain$^{+}$' referred to nonphysical phenomenal pain, a property

[40]Incidentally, this thesis is perfectly compatible with physicalism, even though, if true, it would render the Zombie Refutation in its present form ineffective.

[41]Wittgenstein's private language argument, if sound, would show that it is not possible to refer to an inner state (be that a brain state or a nonphysical phenomenal state) by a concept that does not employ external criteria for its application; so it would show that there could not be a concept like the concept 'pain$^{+}$'. However, this would not count in favor of the Conceivability Arguments, since, as I have pointed out earlier, these same considerations would make zombies inconceivable.

alien to the zombie-word; second, if 'pain$^{+}$' referred to nothing.[42] Either of these scenarios would be damaging to my argument: if zombie-Jackson's term 'pain$^{+}$' referred to pain, then premise (2*) would be false, since all phenomenal$^{+}$ statements would be false in the zombie-world. If, on the other hand, 'pain$^{+}$' didn=t refer to anything, then premise (2*) would be meaningless. Either way, my reductio would not go through. Let=s look at these scenarios one by one.

On the first scenario, the term 'pain' and the term 'pain$^{+}$' have the same meanings. This is not Chalmers=s view: he thinks that the term 'pain' and the term 'pain$^{+}$' have different meanings. He argues like this. In spectrum inverted physical twins, the meanings of phenomenal terms differ. Since, the twins being physically identical, the difference must be due to acquaintance with different phenomenal properties, the zombie's term has to be different from both since the zombie is not acquainted with any phenomenal properties (Chalmers 1996, 207-8). This applies equally in the case of pain or any other phenomenal property.

But these considerations aside, we can see that this scenario is very implausible. By assumption, phenomenal properties are alien to his world. It is quite implausible to assume then that when Jackson says 'that feels good', referring to the phenomenal feels produced by a back-rub, zombie-Jackson also refers to a phenomenal feel, even though there is none in his world. Of course, I am not saying one can never have terms that lack actual reference. The term 'winged horse', for example, has reference. All I am claiming is that, in the particular case of *phenomenal$^{+}$ terms*, like the term 'pain$^{+}$', the reference could not be nonphysical qualia.

'Pain$^{+}$', like 'pain', is a simple term; its reference is not fixed via a description. Let me engage in a little digression here. Not everybody agrees that 'pain' is a simple term. Georges Rey (1988) suggests that even if the reference of 'pain' is not fixed descrip-

[42]Strictly speaking, there is another way in which *Assumption 2* could turn out to be false: if 'pain$^{+}$' referred to a physical state that could not be instantiated by zombies. But that is a very unlikely scenario; I am going to deal with it I when give support to *Assumption 2.*

tively, there is a *descriptive element* to the concept, one that entails that the concept cannot refer to anything physical. It would follow then that *pain*$^{+}$ could not refer to anything physical either. I do not think that our concept has this descriptive commitment; and moreover, I do not think that Jackson or Chalmers thinks that either. Their argument depends on the assumption that we have epistemic warrant to attribute phenomenal states to ourselves on introspective evidence. If our very term 'pain' entailed that its reference is nonphysical, one need not bother with the Conceivability Argument; the truth of dualism would follow from the claim that we have introspective evidence for the occurrence of phenomenal states. But dualism could not be so cheap. On this construal of the term 'pain' we would have to give up our commitment to introspective warrant in attributing phenomenal states. But our assumption that our introspective evidence gives us very strong justification to attribute phenomenal states to ourselves is much more bound up with our term 'pain' than with any claim about its reference being nonphysical. Consequently, it is much more reasonable to hold that the term 'pain' does not have any such descriptive element.

If the term 'pain' is a simple term, what could make it the case that it refers to a nonphysical property? None of the possibilities one can think of would give the result the objector has in mind. It is unlikely that on an interpretationist account zombies could come out referring with their term 'pain$^{+}$' to nonphysical properties alien to their world, as it would render most of their phenomenal$^{+}$ statements false.[43] There are no suitable causal, counterfactual, or lawful relations between nonphysical phenomenal pain and the term 'pain$^{+}$' either. Since there are no pains in the zombie-world, there could not be any causal relations between pain and the term 'pain$^{+}$', as such relations would require the existence of laws connecting physical and nonphysical entities in the zombie-world; but by stipulation, there are no such laws there. The case is the same with counterfactual relations. Another way for reference to be fixed in some direct, nondescriptional manner is

[43]Most, but not all; when one of them says 'I am not in pain$^{+}$,' she would speak the truth.

for it to be fixed by a relation of *acquaintance*. Chalmers (1996, 197) claims that in the case of phenomenal concepts, reference is constituted by *acquaintance* with the referent, where acquaintance is not to be cashed out in terms of causal, counterfactual, or lawful relations. However, this would not help in the zombie-case: zombie-Jackson is just not acquainted with phenomenal experiences in any sense of the word. As the above options exhaust the existing possibilities, we can conclude that the zombie=s simple term 'pain$^{+}$' could not refer to a nonphysical property.

The other objection to *Assumption 2* was that even if the zombie has intentional states in general, his term 'pain$^{+}$' in fact does not refer to anything. In my view, this is wrong. This position has the counterintuitive consequence that all of the zombie's phenomenal$^{+}$ talk lacks truth value. This would be a very uncharitable interpretation of zombies: it would imply that zombies are massively deluded about their mental life. Zombies not only seem to use phenomenal$^{+}$ terms to give reports about their inner states, they also seem to use them to give explanations of each other=s behavior, just as we give explanations of each other=s behavior in phenomenal terms. So, for example, zombie-Jackson=s friend apparently explains why zombie-Jackson takes an aspirin by referring to his headache$^{+}$. Also, their phenomenal$^{+}$ utterances and nonphenomenal$^{+}$ utterances have intelligible connections—zombies say, for example, 'when I had a tooth-ache$^{+}$ last time, I went to the dentist,' etc. These explanations and reports seem to be accurate, since whenever, for example, zombie-Jackson says that he is in pain$^{+}$, he is in a brain state or functional state that is reliably correlated with his term 'pain$^{+}$' (the same brain state or functional state that is reliably correlated with Jackson=s term 'pain'). The natural candidate for the reference of zombie-Jackson=s term 'pain$^{+}$' is this very brain state or functional state.[44] This means that whenever Jackson's statement 'I am in pain' is true, zombie-Jackson's statement 'I am in pain$^{+}$' will be true as well, being about a brain or functional state he is in.

[44]In fact, Shoemaker (1998) has similarly argued that zombies will refer to a brain or functional state by their phenomenal$^{+}$ concepts. He uses the point to a different effect, however; he argues for the view that our phenomenal concepts also refer to physical states, since we are physically identical to our zombie-twins.

The plausibility of this claim might be obscured by the fact that although zombie-Jackson=s statement, for example, 'I am in pain$^{+}$', attributes some brain or functional state to himself, he of course does not conceive of it in this way, that is, he does not think of this state *qua* brain or functional state.

The adherent of the Conceivability Arguments might object at this point that the problem with the alleged term 'pain$^{+}$' is not simply that it lacks reference; the problem is that it does not express a legitimate concept to begin with. As Frege pointed out, distinct concepts must have distinct senses or modes of presentation. Following Frege, White (1986) assumes that modes of presentation have two roles to play simultaneously.[45] On the one hand, they determine reference. On the other hand, they individuate concepts. According to this theory, if the same mode of presentation is associated with two (co-referring) concepts, it must be knowable a priori that these concepts co-refer. No two concepts, where the concepts lack the appropriate a priori links, can have the same reference in all possible worlds. This is apparently because modes of presentations are *properties* of the referent through which the subject grasps the referent.

On this view, zombie-Jackson=s alleged concept 'pain$^{+}$' is not a legitimate one. One could not refer directly to a brain state in the way I have claimed zombie-Jackson must, since that would violate the above principle about modes of presentation. I am assuming that 'pain$^{+}$' and, for example, 'pyramidal cell activity' refer to the same state (a brain state), *via* the same property, since both of these concepts have essential modes of presentation or reference fixers, yet the possessor of these concepts would not be able to know a priori that they co-refer.

Of course, on my view, concept individuation, and so a priori knowledge of co-reference, is more fine-grained than reference fixation. Even if we accept that modes of presentation involve properties of the referent, these properties being what determines the reference of the concept, we might want to deny that these properties exhaust all there is

[45]A very similar argument was formulated by Smart (1959). He introduced his 'topic neutral analyses' of mental terms in response to this argument.

to modes of presentation. Different concepts might employ the same property to provide different routes to the same referent. In the case of the concept 'pain$^{+}$', the same property (for example, pyramidal cell activity) is deployed directly to pick out the referent, whereas in the case of the concept 'pyramidal cell activity', this property is deployed in the way characteristic of scientific terms.

And since this picture of how concepts work is clearly conceivable—that is, there is no incoherence in the idea of concepts that refer directly to brain states in the way I claimed 'pain$^{+}$' does—the burden of proof is clearly on the adherent of the Conceivability Arguments to show, rather than just declare, that such concepts are illegitimate.

Finally, one could object to *Assumption 3*, the assumption that a prioricity supervenes on conceptual role. Here is my defense of it: If a prioricity did not so supervene, then it would be possible that sometimes we cannot tell, even in principle, after a lot of thinking, and doing many thought experiments, of an a priori truth that it is true. If a prioricity did not supervene on actual and potential inferential relations, then we could not claim any special access to a priori truths—a paradoxical situation. Moreover, this would undermine whatever certainty we have in premise (3), the claim that for any true phenomenal statement Q, K -> Q is not a conceptual truth. Denying *Assumption 3* would undermine the Conceivability Arguments by making premise 3 highly contentious.

I would like now to return to *Assumption 1*, the claim that zombies have intentional states, and the objection to it I raised earlier. The objection was that (phenomenal) consciousness is essential to intentionality, and since the zombies, by assumption, do not have phenomenal states, they cannot have thoughts either. Even if it were true that phenomenal consciousness is necessary for intentionality, that would not damage my argument. My argument can be run in a way that would make the objection irrelevant.

In fact, zombie-worlds are introduced only for expository convenience. They are not essential to refute Jackson and Chalmers' argument. My argument against them only presupposes that there is nothing incoherent about the idea of referring to a brain state directly, without the mediation of any physical, functional, or abstract concept, and even

without the mediation of a phenomenal feel figuring as mode of presentation or reference fixer. This will allow me to construct an analogue of the Zombie Refutation that will prove the Conceivability Arguments unsound even if phenomenal consciousness is essential to intentionality.

One way to do this is to consider a world where there are *partial zombies*. If Jackson and Chalmers are right that qualia are nonphysical, then there is a world that is a physical duplicate of our world, but in which there are creatures that have only some of our phenomenal experiences. These creatures will act and talk like us—moreover, they will feel pleasure whenever we do—but they will feel no pain at all (even though they will claim they are 'in pain$^{+}$' whenever we claim we are in pain). Since they do have phenomenal states, and, we might even stipulate, all of their intentional states are accompanied by phenomenal consciousness, there is no reason to deny that they have intentional states. However, on considerations discussed in reply to earlier objections, the most natural thing to say is that their term 'pain$^{+}$' refers to a brain state.

There is another way to make the point in a slightly different way. I submit that the following scenario is at least conceivable—and so, on Jackson and Chalmers=s view, possible. Imagine a world where there are creatures in many respects like us. They have the same physical and mental constitution we have, except that there are some among them that are capable of forming concepts we are not capable of; let us call these people yogis. The yogis are capable of directly detecting certain states of their brains, even though they do not conceive of these states *as* brain states. In some ways, these yogi-concepts will work the way our phenomenal concepts work; they are applied to their referents directly, without the mediation of any physical, functional, or abstract concept. What is peculiar to them is that in the case of the yogi-concepts reference is not even mediated by a phenomenal feel. The yogis will notice that they are capable of detecting *some* inner state of theirs, even though they do not have any idea how they are doing it. Let us call one of the brain states that they can detect in this way state A, and let us suppose that they use the term 'flurg' to refer directly to state A.

Yogis can formulate a variant of the Conceivability Argument. There are true statements in their world involving the concept 'flurg'—for example, 'flurg occurred at t'. These statements will not be derivable from the full fundamental (physical, or if dualism is true, physical *cum* phenomenal) description of their world, since yogis apply their concept 'flurg' directly to brain state A; just like phenomenal concepts, the concept 'flurg' lacks conceptual links to physical, functional, and behavioral concepts sufficient to ground the a prioricity of 'K -> flurg occurred at t'. Yogis then can use the *A Priori Entailment Thesis* to argue that there is a possible world exactly like theirs physically and phenomenally, but where no, as they say, 'flurgs' occur. But such a world is impossible, since, by stipulation, the term 'flurg' refers to a state of their brain. The yogi=s argument is unsound. But among its premises the only contentious one is the *A Priori Entailment Thesis*.

This argument has the advantage of making the same point as the Zombie Refutation, only making it even clearer that the Conceivability Arguments arise not out of any feature specific to phenomenal consciousness, but rather because of a certain peculiarity of our phenomenal *concepts*, a peculiarity that can conceivably be shared by concepts undisputably referring to physical states.

To sum up, even if the objection that phenomenal consciousness is essential for intentionality were sound, it would not succeed in disarming my refutation of the Conceivability Argument. The Zombie Refutation, and its analogues, the partial-Zombie Refutation, and the Yogi Refutation, show that there is something wrong with the Conceivability Argument. It is plausible even on the Zombie Refutation that the premise that has to be given up is the *A Priori Entailment Thesis*; and on the Yogi Refutation this conclusion is inevitable. My arguments then not only show that the conceivability arguments fail. They also prove that Jackson and Chalmers's principle linking conceivability and possibility is false, and so they prove that hopes for grounding all necessities in conceptual and em-

pirical truths were ill founded.[46] Moreover, they help diagnose where things went wrong. The Yogi argument has the advantage over the Zombie Refutation of making it even clearer that the conceivability of zombies arises not out of any feature specific to phenomenal consciousness, but rather because of a certain peculiarity of our phenomenal *concepts*. This peculiarity, that is, referring to a state *directly*, can plausibly be shared by concepts undisputably referring to physical states, and so with regard to these concepts the *A Priori Entailment Thesis* is inapplicable.

## 4 The Aftermath

We have seen that the Conceivability Arguments against physicalism are unsuccessful. In fact, even Jackson, one of the most forceful original proponents of the argument, now thinks that there must be something wrong with it. He thinks that for a dualist, epiphenomenalism is the most reasonable position, given the plausibility of the causal closure of physics. But epiphenomenalism is more implausible than any of the premises are plausible, except the premise claiming that phenomenal states exist. Jackson says that there must be a reply to the Conceivability Arguments, although one cannot quite say what. He calls this the 'There must be a reply' reply (Jackson 1996, 134-35).

With the Zombie Refutation and its companion arguments, we can actually do better. The arguments actually showed where the antiphysicalist went wrong. However, the physicalist, if she wants to make her position attractive, must have an answer to two questions. One is the question of what explains the physicalistic supervenience claims captured in the *Entailment Thesis*:

[46]It is arguable that one can save the spirit, while rejecting the letter, of the *A Priori Entailment Thesis*. In my view, the *A Priori Entailment Thesis* might be correct about all truths except phenomenal truths. This has to do with the special nature of phenomenal concepts. So metaphysical necessity might be reducible to conceptual truth *cum* empirical truths in all cases except cases involving phenomenal concepts (plus the cases mentioned in note 29), and the exceptions themselves might be covered by principles that make modality sufficiently 'un-mysterious'. How exactly to deal with the exceptions, however, is beyond the scope of this paper.

(E) For any true statement T, □(K -> T).

The explanation that Jackson puts forward of why E holds is that all instances of E are conceptual truths. He thinks that the reduction of higher-level concepts to lower-level concepts has to be *perspicuous.* This, however, is unwarranted. The only assumption needed to explain E is that *metaphysical* reductionism is true; that is, the only explanation needed is the assumption that there is some appropriate *metaphysical* relationship (identity, or the realization relation, or perhaps some other, yet unknown relationship) between the referents of higher-level and basic physical concepts. There is no reason such a relation could not hold between physical and phenomenal properties, even in the absence of conceptual connections that would make this relationship perspicuous.

But now, the physicalist also owes an explanation of why phenomenal statements appear to be so different from other higher-level statements in their connections with lower-level discourse. Many of us are convinced (partly by the Conceivability Arguments) that there is something special about phenomenal statements. It seems right that it is not conceivable, after all the physical truths are in, that water is not $H_2O$. But it is still conceivable that any phenomenal statement is false, no matter how much physical information we have. And the question of the explanatory gap remains as well.

However, there is no mystery about all this. The explanation of this should be rather obvious by now. Physicalists who adopt a direct recognitional account of phenomenal concepts will not be in the business of trying to close the gap, or explaining away the conceivability of zombies, since, on this account of phenomenal concepts, it is *to be expected* for a physicalist that there will be an explanatory gap, and that zombies are conceivable. In the Yogi Refutation I have constructed a concept that refers to a physical state even though the fact that it does so is not derivable a priori from the full physical description of the world. There is even less a priori reason to rule out the possibility that something like this is the case with phenomenal terms. On this account, we get the fol-

lowing picture.[47] Phenomenal concepts are direct recognitional concepts and they employ as their reference fixer the very state they are denoting: the itchy feeling of an itch serves to fix the reference of the phenomenal concept *itch*. Phenomenal concepts, on the other hand, *refer to the very same property as some neurophysiological (ultimately, microphysical) concept*: assuming that an itch just *is* a certain brain/functional state, there will be an appropriate neurophysiological/functional concept whose reference fixer will involve the same property (a certain neurophysiological/functional property that is identical to an itch); only the reference fixer is deployed in the way characteristic of scientific terms. A phenomenal concept and a concept of microphysics, each of which picks out its referent through an essential reference fixer (say, some neurophysiological property), could then refer to the same property, even in the absence of the kind of conceptual connections required by the *A Priori Entailment Thesis*.

But what about the persistent intuition that, despite every argument in favor of it, physicalism *just can't be true*? I think that it can be explained by the intuitive pull of the Transparency Thesis, roughly, the thesis that if we have two concepts, both of which refer via an essential reference fixer, then we *must* be able to tell whether they co-refer. After all, we have an insight into the *nature* of their referent through their reference fixers; so how could we be wrong about our judgment (as it is in the phenomenal/neurophysiological case) that they do not co-refer?[48] But in the light of the picture of phenomenal concepts given above, this intuition is shown to be misplaced.

One might object that this explanation does not do justice to our Yogi thought experiment. In the Yogi Refutation I have hypothesized that there could be beings that possess concepts directly referring to (nonphenomenal) physical states. Given my refutation of the Conceivability Arguments, I cannot claim, merely on the basis of their conceivability, that they are possible. But I see no reason why they would not be. Yogis can make

[47]The following originates in Loar 1997. See also Sturgeon 1994.

[48]This is a close relative of the theory of concepts I attributed to White (1986) earlier, where I was considering the viability of the concept 'pain$^{+}$'.

statements that are true in their world even though these statements are not derivable a priori from the full fundamental description of their world.

Yet it is plausible to speculate that yogis would not be inescapably drawn to dualism. But should not they be, given our claim that a belief in the Transparency Thesis is enough to explain antiphysicalist intuitions? Presumably, yogis are just as attracted by the Transparency Thesis as ordinary humans are. But there is no contradiction here. The yogi can be attracted by the Transparency Thesis, and still not be drawn to dualism, just on the basis of her special conceptual repertoire. There is a difference between us *vis-à-vis* phenomenal concepts and the yogi *vis-à-vis* the yogi concepts; the yogi, as opposed to us, does not have a temptation to think that she has direct insight into the nature of what her concept *flurg* refers to. In a sense, she does not have a handle on the concept; the reference fixer of her concept might be an essential property of the referent, but she does not have access to it, the way we have access to the phenomenal reference fixers of our phenomenal concepts.

We started with the question of what the Conceivability Arguments can teach us about the mind-body problem. On the present picture of phenomenal concepts, the conceivability of zombies is a symptom of the unique role phenomenal concepts play in our conceptual repertoire, but it is not a guide to their possibility. This is not the lesson intended; but, all the same, it is an important one.

**Bibliography**

Balog, Katalin. 1998. Conceivability Arguments. Ph.D. diss., Rutgers University.

Chalmers, David. 1996. *The Conscious Mind*. New York: Oxford University Press.

Cottingham, John, Robert Stoothoff, and Dugald Murdoch, eds. 1984. *The Philosophical Writings of Descartes*. Cambridge: Cambridge University Press.

Dretske, Fred. 1988. Explaining Behavior: Reasons in a World of Causes. Cambridge: MIT Press.

Fodor, Jerry. 1990. *A Theory of Content and Other Essays*. Cambridge: MIT Press.

Fodor, Jerry. 1997. *Locke Lectures*.

Jackson, Frank. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32:127-36. Reprinted in *Mind and Cognition,* ed. William Lycan, 469-78. Oxford: Blackwell, 1990.

Jackson, Frank. 1993. Armchair Metaphysics. In *Philosophy in Mind*, ed. M. Michael and John O'Leary-Hawthorne. Dordrecht: Kluwer.

Jackson, Frank. 1998. *From Metaphysics to Ethics*. New York: Oxford University Press.

Jackson, Frank, and David Braddon-Mitchell. 1996. *Philosophy of Mind and Cognition*. Cambridge: Blackwell.

Kripke, Saul. 1972. *Naming and Necessity*. Cambridge: Harvard University Press.

Levine, Joseph. 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64:354-61.

Levine, Joseph. 1993. On Leaving Out What It=s Like. In *Consciousness: Psychological and Philosophical Essays,* ed. Martin Davies and Glyn W. Humphreys, 121-136. Oxford: Blackwell.

Levine, Joseph. 1998. Conceivability and the Metaphysics of Mind. *Noûs* 32:449-80.

Lewis, David. 1966. An Argument for the Identity Theory. *Journal of Philosophy* 63: 17-25.

Lewis, David. 1983. New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61:343-77.

Loar, Brian. 1990. Phenomenal States. *Philosophical Perspectives* 4:81-108.

Loar, Brian. 1997. Phenomenal States. In *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere, 597-617. Cambridge: MIT Press. (Revised version of Loar 1990).

Loewer, Barry. 1995. An Argument for Strong Supervenience. In *Supervenience*, ed. Elias E. Savellos and Ümit D. Yalcin. Cambridge: Cambridge University Press.

Millikan, Ruth. 1989. Biosemantics. *Journal of Philosophy* 86: 281-97.

Nagel, Thomas. 1974. What Is It Like to Be a Bat? *Philosophical Review* 83:435-50.

Papineau, David. 1993. *Philosophical Naturalism*. Oxford: Blackwell.

Quine, Willard O. 1960. *Word and Object*. Cambridge: MIT Press.

Rey, Georges. 1988. A Question about Consciousness. In *Perspectives on Mind*, ed. H. Otto and J. Tueidio. Dordrecht: Kluwer.

Robinson, Howard. 1993. The Anti-materialist Strategy and the Knowledge Argument. In *Objections to Physicalism,* ed. Howard Robinson. Oxford: Oxford University Press.

Ryle, Gilbert. 1949. *The Concept of Mind.* London: Hutchinson.

Searle, John. 1992. *The Rediscovery of the Mind.* Cambridge: MIT Press.

Shoemaker, Sydney. 1981. Absent Qualia are Impossible. *Philosophical Review* 90:581-99.

Shoemaker, Sydney. 1998. Commentary in Symposium on Chalmers= *The Conscious Mind.* Forthcoming in *Philosophy and Phenomenological Research*.

Sturgeon, Scott. 1994. The Epistemic View of Subjectivity. *Journal of Philosophy* 91:221-36.

White, Stephen. 1986. Curse of the Qualia. *Synthese* 68:333-68.

Witmer, Gene. 1997. Demanding Physicalism. Ph.D. diss., Rutgers University.